# The American Economic Review



## ARTICLES





MARCH 1976

# THE AMERICAN ECONOMIC ASSOCIATION

Founded in 1885

• Published at George Banta Co., Inc., Menasha, Wisconsin.

• THE AMERICAN ECONOMIC REVIEW, including four quarterly numbers, the *Proceedings* of the annual meetings, the *Directory*, and Supplements, is published by the American Economic Association and is sent to all members five times a year, in March, May, June, September, and December.

Association membership dues for 1976, which include a subscription to both The *American Economic Review* and the *Journal of Economic Literature*, are as follows:

$25.00 for regular members with rank of assistant professor or lower, or with annual income of $12,000 or less;

$30.00 for regular members with rank of associate professor, or with annual income of $12,000 to $20,000;

$35.00 for regular members with rank of full professor, or with annual income above $20,000;

$12.50 for junior members (registered students). Certification must be submitted yearly.

Subscriptions (libraries, institutions, or firms) are $37.50 a year. Only subscriptions to both publications will be accepted. Single copies of either journal may be purchased from the Secretary's office, Nashville, Tennessee.

For foreign postage, add: PUAS countries, $2.10; all other countries and Canada, $2.50.

• Correspondence relating to the *Papers and Proceedings*, the *Directory*, advertising, permission to quote, business matters, subscriptions, membership and changes of address should be sent to the Secretary, C. Elton Hinshaw, 1313 21st Avenue South, Nashville, Tennessee 37212. Change of address must reach the Secretary at least six (6) weeks prior to the month of publication. The Association's publications are mailed second class.

• Second-class postage paid at Nashville, Tennessee and at additional mailing offices. Printed in U.S.A.

# THE AMERICAN ECONOMIC REV

March 1976

VOLUME 66, NUMBER 1






• Manuscripts and editorial correspondence relating to the regular quarterly issue of this REVIEW should be addressed to George H. Borts, Managing Editor of THE AMERICAN ECONOMIC REVIEW, Brown University, Providence, R.I. 02912. Manuscripts should be submitted in duplicate and in acceptable form and should be no longer than 50 pages of double-spaced typescript. A submission fee must accompany each manuscript: $15 for members; $30 for nonmembers. *Style Instructions* for guidance in preparing manuscripts will be provided upon request to the editor.

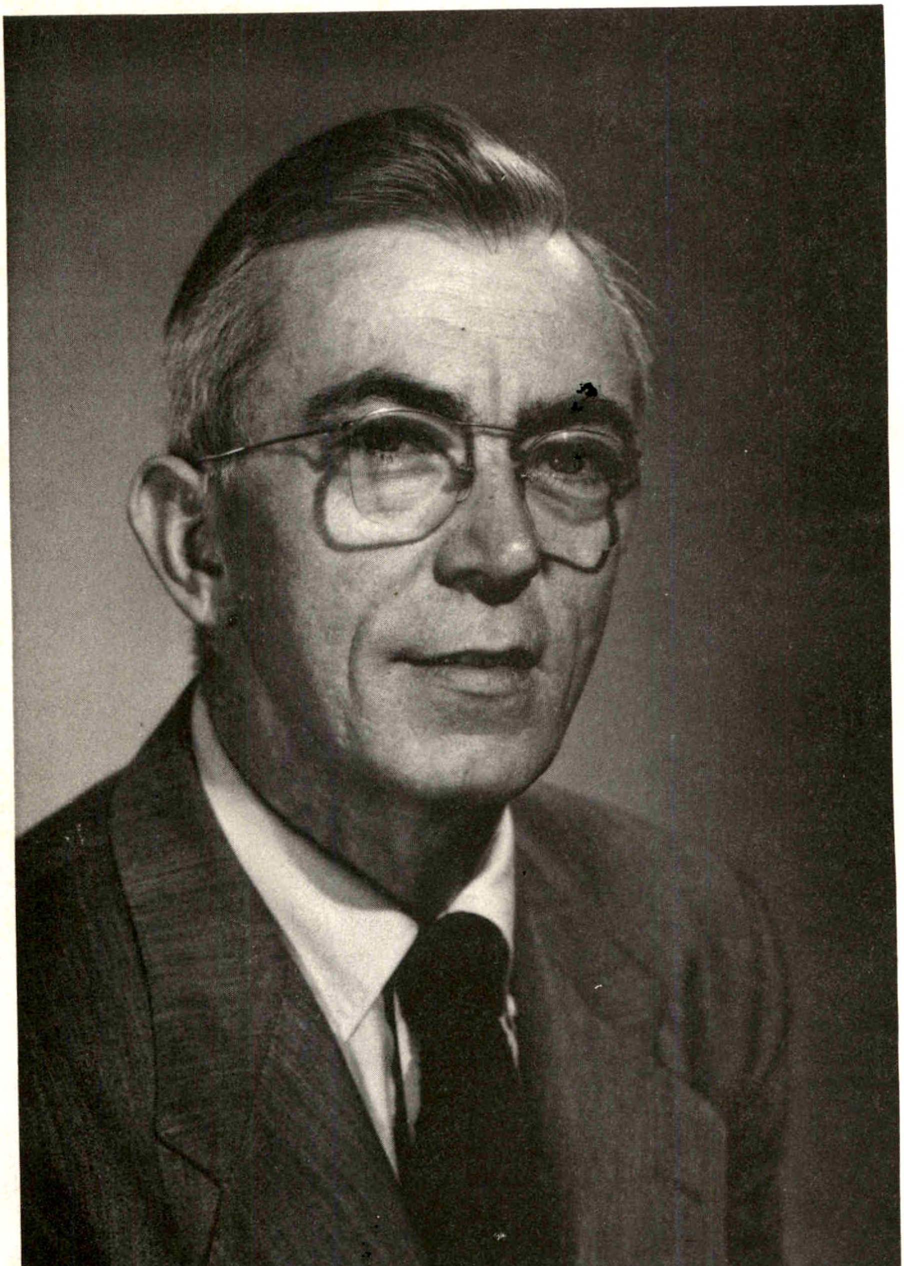
• No responsibility for the views expressed by authors in this REVIEW is assumed by the editors or the publishers, The American Economic Association.



**Articles**

ter Papers

Number 77 of a series of photographs of past presidents of the Association



R A Gordon

# Rigor and Relevance in a Changing Institutional Setting

By Robert Aaron Gordon*

The title of this paper summarizes the two-fold theme to which I want to address myself this evening. First, the mainstream of economic theory sacrifices far too much relevance in its insistent pursuit of ever increasing rigor. And, second, we economists pay too little attention to the *changing* institutional environment that conditions economic behavior. We do not often enough reexamine our basic postulates in light of changes in this environment, and, perhaps more important, we shy away from the big questions about how and why the institutional structure is changing—and where it is taking us.[1]

## I

Economists pride themselves on belonging to the most "scientific" of the social sciences. The justification for this contention lies in the growing resemblance between the nature of the analytical tools used in economics and in the natural sciences—above all, the increasing use of mathematical tools in theoretical analysis and the development of sophisticated mathematical and statistical techniques in empirical work. Today, mathematically formulated economic theory, the development of econometric techniques, and the sophisticated application of econometric methods to the "testing of hypotheses" in a variety of applied fields constitute the core of the science of economics.[2]

What is science? One brief definition runs: "A systematic knowledge of the physical or material world." Most definitions emphasize the two elements in this definition: (1) "systematic knowledge" about (2) the real world. Without pushing this definitional question to its metaphysical limits, I merely want to suggest that if economics is to be a science, it must not only develop analytical tools but must also apply them to a world that is now observable or that can be made observable through improved methods of observation and measurement. Or in the words of the Hungarian mathematical economist János Kornai, "In the real sciences, the criterion is *not* whether the proposition is logically true and tautologically deducible from earlier assumptions. The criterion of 'truth' is, whether or not the proposition corresponds to reality" (p. 9, his italics). (Compare Wassily Leontief 1966, p. 23.)

One of our most distinguished historians of economic thought, George Stigler, has stated that: "The dominant influence upon the working range of economic theorists is the set of internal values and pressures of

* Presidential address delivered at the eighty-eighth meeting of the American Economic Association, Dallas, Texas, December 29, 1975. My thanks for helpful suggestions go to Moses Abramovitz, G. L. Bach, John Gurley, and Benjamin Ward. And being uninhibited by virtue of their membership in the same family, Margaret S., David M., and Robert J. Gordon not only offered suggestions but forced me to rewrite substantial sections through successive drafts. Finally, I want to thank James Hodder and Carl Walsh for their assistance, which was made possible by the Institute of Industrial Relations, University of California, Berkeley.

[1] A number of economists have anticipated various of the points made in this paper, and some are cited later. In addition, I should mention E. H. Phelps Brown and G. D. N. Worswick. And there are others, several of whom were cited by Walter Heller in his presidential address last year.

[2] I trust that noneconometric applied economists, including Nobel prize winner Simon Kuznets, will forgive me for thus excluding them from the "core" of economics.

the discipline. The subjects of study are posed by the unfolding course of scientific developments" (p. 22). He goes on to add: "This is not to say that the environment is without influence. . . . " But, he continues, "whether a fact or development is significant depends primarily on its relevance to current economic theory" (p. 23). (Compare Tjalling Koopmans, p. 170.) What a curious relating of rigor to relevance! Whether the real world matters depends presumably on "its relevance to current economic theory." Many if not most of today's economic theorists seem to agree with this ordering of priorities.[3]

To what aspects of the observable world does economics apply its analytical tools? According to the familiar definition in the *International Encyclopedia of the Social Sciences*, which we owe originally to Lionel Robbins, economics "is the study of the allocation of scarce resources among unlimited and competing uses." To this definition of microeconomics the *Encyclopedia* then rather weakly adds macroeconomics, which is defined as the study of money, the general price level, and the level of output and employment.

Let us consider the microeconomic part of this definition. Presumably the reference is to real resources, used to produce observable goods and services, that are exchanged and consumed by real people living in the kind of world we see around us. Of course, some degree of abstraction is necessary if useful generalizations are to be reached. Here the economic theorist quickly runs up against a dilemma. Shall he seek to make his analysis ever more rigorous, regardless of the possibly diminishing relevance of his conclusions to the observed world? Or shall he sacrifice elegant refinement for somewhat cruder analysis that may lead to testable results? I suppose the reply can be given that we need to do both and that we need specialists in each of these approaches. But certainly those who construct the analytical apparatus should pay more attention than many of them now do to the substantive problems with which economists are presumably concerned.[4]

In speaking of how well current microeconomic theory combines rigor and relevance, I should distinguish among the different but interrelated problem areas with which this part of economic theory concerns itself. Micro theory addresses itself primarily to three related topics: (1) the conditions necessary for, and the means of actually achieving, an optimum allocation of resources *within* decision-making units, both firms and households, under given assumptions as to the criteria of optimization; (2) again under given assumptions, the conditions necessary for the existence of a general (or partial) equilibrium among all (or some) of these decision-making units, including the determination of the uniqueness and stability of such an equilibrium; and (3) the conditions required for the achievement of an economic optimum from a broad, social point of view.

Some success has been achieved in blending rigor and relevance in the theory of the single decision-making unit. Certainly, as this year's Nobel prize in economics attests, considerable progress has been achieved in dealing with production planning in the individual firm or establishment (or government department) for which mathematical tools, including programming and the whole range of activity analysis, have proved to be useful. To some extent, how-

[3] This has led Leontief to expostulate: "Seldom, in modern positive science, has so elaborate a theoretical structure been erected on so narrow and shallow a factual foundation" (1966, p. 33). Compare also Nicholas Kaldor, p. 1240, and Martin Shubik (1970).

[4] For a careful and judicious defense of theoretical models, see Koopman's second essay. He also adds: "Precision and rigor in the statement of premises and proofs can be expected to have a sobering effect on our beliefs about the reach of the propositions we have developed" (p. 147).

ever, the success achieved here makes this part of economics resemble more a branch of engineering than a social science. And at the level of the individual firm some progress continues to be made in empirical studies of production and cost functions, the determinants of the demand for inputs, the transmission of technical change, and related topics. I do not wish to minimize the value of this work. But we should not ignore the extent to which rigorous formulations of the theory of the firm have had to be relaxed in order to obtain useful results in empirical work. Nor, I might add, should we forget the extent to which conventional theory ignores how and why work is organized within the firm and establishment in the way that it is, what may be called the "social relations" of the production process.

Some success in blending rigor and relevance has also been achieved in the field of household behavior—ranging from studies of the determinants of consumers' demand to recent work on human capital, the behavior of labor markets, the economics of crime, and the like. I must confess to some skepticism, however, about the relevance of the economic models of household behavior recently developed by Gary Becker and his followers—what has been referred to as "the new home economics." Granted that much useful work has been done in this area. Nonetheless there is a lamentable tendency among scholars in this field to rely upon a caricature of human beings who continuously and consciously balance costs and benefits at the margin, whether in deciding on another year of schooling, whether and when to marry or be divorced, how many children to have and when, or whether and when to commit a crime. And after a substantial amount of intensive research, the human capital approach still leaves unexplained a significant part of the differences in personal incomes.

In the second area mentioned, particularly general equilibrium theory, it seems to me that relevance has been largely absent in the recent literature. To find much relevance at the theoretical level, and I refer here only to the theoretical literature, we must go back to the partial-equilibrium analysis of Alfred Marshall and his followers. Walras, Pareto, and their successors—with their assumptions of atomistic competition, perfectly flexible prices, costless information, and limitless futures markets—have contributed little to relevance in their steadfast pursuit of rigor. Although a step in the right direction, recent attempts to develop a pure theory of disequilibrium are subject to much the same criticism.

Rigor and relevance have been successfully blended in input-output analysis, although largely at the expense of ignoring the price sensitivity of input-output coefficients. Here again, the emphasis is on engineering-type relationships, but in this case the entire economy is the object of study.

Micro-economic analysis, as I have noted, concerns itself with "the allocation of scarce resources among unlimited and competing uses." As economists we are concerned with how resources may be allocated *efficiently*, and we are prepared to provide the layman and the policymaker with a rigorous definition of efficiency. But "efficient allocation" for whose benefit? To me it has always been startling that the accepted province of micro-economic theory has little room for the personal distribution of income—and virtually none for the personal distribution of wealth. Of course, we speak of "distribution," but by this we mean either factor prices or total factor shares.

Why wealth and labor services, the latter carrying widely different market-determined prices, are distributed among different human beings in the way they actually are is a question that the non-Marxian

economic theorist seldom asks. An exception should be made for the relevant literature in the field of labor economics. Similarly, some exception should be made for the recent literature on human capital, but at best it is only a partial exception.[5] The mathematically inclined general theorist continues to show little interest in the determinants of the personal distribution of income and wealth.[6] It has been said by George Stigler that: "The problem of personal income distribution will eventually receive much theoretical attention, since it is a problem of all economies and all times" (p. 22). (Compare Becker, p. 135.) But if the problem is so important, why only "eventually"? It is not surprising that many younger economists are seeking a radical alternative for the neoclassical framework, more or less along Marxian lines. And it is also this concern with the unequal personal distribution of opportunity and income which has led to the development of models of a dual labor market. These models, with their emphasis on institutional barriers to labor mobility, offer some valuable insights into the operation of contemporary labor markets, and they also raise important questions about some of the assumptions implicit in neoclassical theory.[7]

From the point of view of human welfare—a concept that will not go away no matter how uncomfortable it makes the economic theorist—can we ignore the personal distribution of income? Which is more relevant: a rigorous demonstration as to how resources can be most efficiently allocated under ideal conditions that have never existed, or a much cruder exploration of how wealth and income came to be distributed as they in fact are and what might be done to affect the distribution of income in one way or another? As Alice Rivlin put it in her Ely lecture last year (p. 2), economists "worth their pay" ought to be able to explain the shape of the income distribution and why it is changing or not changing. By this criterion very few of us are worth our pay.[8] To go further, why do we have so little to say about the intergenerational movement among occupational and income classes, about the determinants of the distribution of income relative to those affecting the distribution of wealth, and about the ethnic, social, political, and regional factors affecting the distribution of both wealth and income?

Am I suggesting that economic theory become much more normative than it now is? Of course I am not. "Relevant" and "normative" are not synonyms, and what I am alleging here is that neoclassical economics has failed to be relevant in its refusal to deal with the personal distribution of income and wealth. This refusal stems, I presume, from the fact that, with the analytical tools at hand, the problem has seemed too difficult.

I should add in this connection that neoclassical economics has always had a normative slant. As others have suggested,

[5] In the second edition of *Human Capital*, pp. 94–144, Becker republishes his earlier attempt to utilize the human capital approach to sketch out a theory of personal income distribution. See also Jacob Mincer.

[6] On the various approaches to the study of the distribution of income, see the useful volumes by Martin Bronfenbrenner and Jan Pen. Interestingly, some three-quarters or more of Bronfenbrenner's book is devoted to the functional distribution of income; only one chapter is devoted entirely to "Topics in Personal Income Distribution." Yet one reviewer criticized him, as well as Pen, for devoting too much space "to the distributional influence of institutional forces and to the personal, as against functional, distribution of income" (Charles E. Ferguson, p. 440). Pen's volume contains a useful summary of the more important recent literature on the determinants of the personal distribution of income; and, interestingly, his longest chapter is concerned with "norms and policies" for income redistribution.

[7] Compare Arthur Okun (1973, pp. 237–46) and Michael Wachter and the references there cited. Cairnes and Marshall, of course, did talk about noncompeting groups, but little of this has carried over into the main body of current economic theory.

[8] One who has recently earned his salary is Okun (1975); see also James E. Meade and Alan Blinder. An earlier example of an economist who "bucked the trend" to discuss the personal distribution of income and wealth was Hugh Dalton.

conventional micro-economic theory as it has developed, particularly in the last seventy-five years or so, takes a normative stance by default.[9] Indeed, it takes a normative stance by more than default. It says outright that the primary question to which economists should address themselves is the "optimum" allocation of resources, and it insists on providing a precise definition of optimum. (Compare Ward, pp. 52–54, 90.) Since, as Koopmans points out, competitive equilibrium theory ignores the welfare implications of the personal distribution of income that results, "the term 'optimum' [is] a misnomer" (p. 49). (See also Ward, pp. 197–98.)

My remarks thus far have been directed toward microeconomics. But macro-economic analysis also must face the problem of how optimally to combine rigor and relevance. What we today call macroeconomics grew out of the catastrophe of the Great Depression,[10] and in the early development of macro-economic analysis relevance took precedence over rigor. Today, rigor competes with relevance in macro-economic and monetary theory, and in some lines of development macro and monetary theorists, like many of their colleagues in micro theory, seem to consider relevance to be more or less irrelevant. A good example has been the elaboration of so-called growth theory which now seems to be losing some of its earlier popularity. Apparently, there has come to be a growing recognition that the considerable efforts spent on growth models have not significantly advanced knowledge beyond the contributions of the original Harrod-Domar and the Solow-Swan neoclassical models. And other intriguing topics have arisen to stimulate the pursuit of rigor at the expense of relevance.

In both micro- and macroeconomics, efforts are sometimes made to extract a drop or two of relevance from exercises in analytical rigor; and conclusions are drawn about the functioning of some aspect of the real world, or policy recommendations are made, on the basis of theoretical exercises which rest on assumptions that fly in the face of the facts. One example is the frequent fitting of the neoclassical investment function, with its built-in assumptions of constant returns and perfect competition, to industries in which these assumptions obviously do not hold. Another example is a good deal of the recent literature seeking to reformulate the micro-economic foundations of macro-economic theory and policy—what Edmund Phelps (1969, 1972) has termed "The New Microeconomics in Inflation and Employment Theory." The theoretical analysis in much of this literature rests on assumptions that also fly in the face of the facts. To cite a few examples: all unemployment is a voluntary activity as part of a search procedure in which workers are continuously equating costs and prospective benefits at the margin; the labor supply is typically taken to be homogeneous with perfect mobility among labor submarkets; so-called structural unemployment is ignored as are the striking differences in unemployment rates among different age, sex, ethnic, and occupational groups; and downward wage flexibility is generally assumed, although some recent attempts have been made to relax this assumption. Another related recent development in which theory proceeds with impeccable logic from unrealistic assumptions to conclusions that contradict the historical record, is the recent work on rational expectations. (Compare Robert J. Gordon.) And, as a final illustration I might cite much of the recent literature on capital theory.

## II

I turn now from the first part of my title to the second—from rigor and relevance to

[9] Marxian critics might add that the normative stance results from ideological bias.

[10] To say this is not to minimize the contributions of Keynes' predecessors, particularly Wicksell.

the fact that we live in a world that is continually changing. And here I want to pose two questions: First, to what extent does the changing institutional environment affect the relevance of the analytical tools that we use and the assumptions that we make about the determinants of individual and group behavior? (Here, of course, I am still raising the question of relevance.) And, second, why do we ask so few questions about why and how the institutional environment has changed in the way that it has, and what are its internal dynamics that will lead it to change in particular ways in the future—not only in the United States but in other countries? A few economists have addressed themselves to this range of questions, notably Karl Marx and Joseph Schumpeter. Among living economists, three who at least raise this range of issues are John Kenneth Galbraith, Gunnar Myrdal, and, in the Marxist tradition, Paul Sweezy.[11]

Certainly the outstanding example of the failure of economic theory to adapt its analytical tools to the changing institutional environment must be the stubborn adherence to the assumption of perfect competition, a concept which has been described as being "as pervasive and fundamental as any in the whole structure of classical and neoclassical economic theory" (Stigler, p. 234).[12] Indeed for a century or so, economists have toiled to make more precise the notion of a perfectly competitive market. Over this same period, of course, the character of actual markets has been changing in many ways. While improvements in transportation and communication have tended to promote competition in expanding markets, the growth of large firms and the spread of industrial concentration have made oligopoly a much more relevant model for industrial markets than the perfectly competitive model which today's theorists insist on using. It is true that sporadic efforts have been made to develop a theory of oligopoly, and for a while high hopes were held for what might be learned from game theory; but no generally accepted theory of oligopoly has yet emerged. At the same time, the emphasis on general equilibrium theory has tended to turn attention away from this egregious departure from perfect competition.[13]

It is true, of course, that in the 1930's, under the stimulus of the pioneering works by Edward Chamberlin and Joan Robinson, we developed a theory of monopolistic or imperfect competition, centering on the notion of product differentiation. And at the same time increased attention began to be paid to the determinants of oligopolistic behavior. At the applied level, the field of industrial organization was born. But while this applied field has continued to thrive, general micro-economic theory and the applied work in this area have largely parted company. General equilibrium theory (and not only this branch of theory) has returned to the assumption of perfect competition. The notion of a sloping demand curve for the individual firm seemingly did not add much to the general theorist's tool box, and the new mathematical economics found it more exciting to pick up the challenge of Walras and Pareto and to turn to general equilibrium analysis and to setting forth the conditions of Pareto optimality. And for this the assumption of perfect competition was convenient if not essential. As William Baumol

[11] There are, of course, a fair number of economists who consider themselves to be in the American institutionalist tradition. They are likely to belong to the Association for Evolutionary Economics, which has its own journal, but thus far they have had little influence on the main trends in theoretical and applied work in economics.

[12] See also Schumpeter (1954, pp. 972 ff).

[13] See, for example, the papers by Paul Joskow and Martin Shubik in last year's *Papers and Proceedings* of the American Economic Association. See also Shubik (1970, p. 415), who bluntly declares: "There is no oligopoly theory."

has noted, "The case of product differentiation has proved particularly resistant to general equilibrium analysis" (p. 45).[14]

And so, as power blocs multiplied in a pluralistic world, as firms grew larger and as conglomerates were added to vertical and horizontal combinations, as advertising expenditures mounted to influence spending out of rising discretionary income, as the problem of externalities became ever more important, and as the role of government in the functioning of markets steadily increased, micro-economic theory largely averted its eyes and became ever more enamored of hypothetical systems of general equilibrium under conditions of perfect competition.[15]

In the meantime, instructors in undergraduate micro theory have resolutely continued to teach their students the essentials of the Chamberlinian partial equilibrium analysis, with downward sloping demand curves facing the individual firm, and laying down the conditions for short-run and long-run equilibrium on an industry basis. That much, at least, undergraduates seem able to absorb, and some factual counterparts can be found for the theoretical analysis. But at the graduate level in many universities, and the more so the more advanced the level of instruction, Walras and Pareto and their successors take over, and the analysis proceeds on the basis of the conditions associated with atomistic competition.

A further example, reflecting another dimension of the growth of large-scale business, involves the conditions under which decisions are made in the large firm.[16] Involved here are a number of issues which have been the subject of some theoretical analysis and a good deal of empirical work in the postwar years, but very little of the results of this work has yet found its way into main corpus of micro-economic theory. I have in mind such questions as the relative roles of profit maximizing and of other criteria in the decision rules used in the large firm, the effects of the separation of ownership and management on these optimization criteria, the ways in which the bureaucratization of decision making affects the manner in which business firms respond to market stimuli, and the effect of the improvements in the gathering and processing of information that have come not only from the revolution in data processing but also from the development of a more scientific approach to decision making.[17] What has been the effect of all this on the pricing decisions of large firms, the manner in which they participate in wage negotiations, or how their investment planning responds to current and prospective developments in product and financial markets? And there are other examples of the same sort. Early work not only in

[14] Kenneth Arrow (1971) has since attempted to fit into general equilibrium analysis the case in which there are both monopolistic and competitive firms. I should say that this exercise was characterized more by rigor than by relevance.

[15] I am speaking here only of the persistence of the assumption of perfect competition in theoretical work, regardless of the changes to which I refer. This is not to deny that useful theoretical work has been done in some of these problem areas.

At this point, I cannot refrain from quoting Arrow: "The prestige status of the purest of pure economic theory has never been higher; and yet there is now, as there has always been, a pervasive skepticism about the descriptive power and normative utility of Walrasian and other varieties of the theory of general competitive equilibrium" (1967, p. 733).

[16] Here I emphasize decision making in the *large* firm. Much of the recent theoretical work on decision making under uncertainty has taken place in an institutional vacuum and is presumably applicable to small as well as to very large decision-making units. On the other hand, what is conventionally referred to as organization theory generally addresses itself to the relatively large organization. For an innovative attempt to develop a theory of economic behavior for the large firm, see Robin Marris (1964). In this context, reference should also be made to the work of Richard Cyert, James March, Herbert Simon, and their students at Carnegie-Mellon.

[17] For a survey of some of the relevant recent literature with an emphasis on the institutional setting for decision making, see Eirik Furubotn and Svetozar Pejovich. The issues raised here are also emphasized by Galbraith.

economics but also in the other social sciences led to the development of what is now called "organization theory," and this is now a field in the better business schools; but not much of this has seeped into the mainstream of economic theory.

Some economic theorists today are beginning to pay serious attention to the beginnings of a theory of information, and increasing attention is being paid to problems of decision making under uncertainty. But here again too little attention tends to be paid to the changing conditions under which information is collected and processed, and the manner in which institutional arrangements affect the way in which the future is viewed and attitudes toward uncertainty change.

There is another and indeed startling respect in which pure micro-economic theory tends to ignore the changing institutional environment. This has to do with the steadily increasing role of government in the functioning of markets. In neoclassical general equilibrium theory, there is no place for any kind of public authority. Of course, we have the field of public finance, which on the theoretical side and at the micro level does borrow from neoclassical theory; and, as Walter Heller reminded us last year, cost-benefit analysis and tax-incidence theory have helped economists to develop a rationale for many types of government decision making. There is also a growing literature on the economics of government decision making in a range of problem areas—from pollution to military spending to housing and education. But the pure micro theorist finds no role for a public authority in his analysis of the determinants of general equilibrium. (See Bent Hansen, p. 92.)

Here we encounter an important difference between micro and macro theory. Much of macro theory (but by no means all) tends to be policy oriented and to have a strong normative orientation. It is concerned both with policy variables and with variables that can be, directly or indirectly, influenced by policy. While there is a good deal of macro theory in which the possible role of government is ignored, much of theoretical macroeconomics does, either explicitly or by implication, leave an important role for government.

Although macro-economic theory does not ignore the role of government, we can still find plenty of examples of the failure of macro theory to reflect the changing institutional environment. One of the most striking examples undoubtedly involves what little we have in the way of a theory of inflation. The problem is most acute with respect to accounting for inflationary trends since World War II.

A number of explanations are currently circulating regarding the tendency toward accelerating inflation in the last decade or more. The purely monetarist one is the simplest and makes no reference to a changing institutional structure. It tends to ignore the sources and nature of the pressures, operating through government, which lead to changes in the supply of money. An increasing number of economists, monetarists and Keynesians, emphasize the existence of a natural rate of unemployment, with the implication that government policies to expand aggregate demand, by pushing unemployment to or below the natural rate, lead to accelerating inflation. But advocates of the natural rate hypothesis have little to say about why the natural rate, if it exists, is what it is today, and how and why it has changed over the years. Much of the work in this area tends to be done in an institutional vacuum, including the recent work on the formation of price expectations.

I think it is fair to say that we still lack a general theory with a significant time dimension of the nature of the inflationary process and how this process is affected by the changing institutional setting. Gunnar

Myrdal is one of those to emphasize that "in modern society the tendency of inflation to become cumulative and to accelerate is rooted in a wide and complex institutional reality" (p. 24), and to urge a broader, institutionally oriented analysis of underlying causes and possible remedies. Another is Peter Wiles in a provocative recent essay in the *Economic Journal.* Among these institutional changes making for more rapid and apparently accelerating inflation are the postwar commitment in all advanced countries to a high level of employment (stronger in Europe than in the United States), the reluctance fully to cover increasing public expenditures by taxation, the growing tendency to link wage and other payments to the consumer price index, the increasing aggressiveness of trade unions (emphasized by Wiles) and other organized groups of income receivers, an apparent weakening of the willingness of employers to resist wage demands (which is closely related to the government's commitment to high employment), and the intensification of inflationary expectations engendered by past inflation. (See Myrdal, pp. 23 ff.) To all of which we can add an international monetary system that permitted huge dollar outflows which became expanded monetary reserves for other countries. This is a familiar list. But how do these different but related inflationary pressures interact? Why have they become stronger? How do we account for the particular rate of acceleration that has occurred? Why does the rate of acceleration seem to be different, and because of what differences in the institutional environment, in different countries? And how have the significant changes in the international system of trade and finance affected these inflationary trends? (See R. J. Gordon.)

I shall merely mention one other example of the failure of economists today to take adequate account in setting up their models of the changing institutional environment. This has to do with the determinants of household behavior, with respect to both the demand for goods and services and the supply of labor services. What has been the effect on household behavior, for example, of advancing levels of education, changes in the ways that news is disseminated, recent trends toward urbanization and suburbanization, or the massive change in the impact of advertising made possible by television?[18] To what extent have these and other developments made households behave more or less in the way that economic theory assumes? What bearing do these and other institutional developments have on the behavior of personal saving in recent years, in this and other countries, and on the way that households respond to recent and prospective inflationary trends?

## III

I turn now to my second question about the changing institutional environment. Why does the central body of economic analysis show so little interest in why and how the institutional setting for economic behavior has changed in the past and is likely to change in the future? The past, of course, is the domain of the historian, but I am not aware that the vast amount of historical research in the past century or more has yet given us an acceptable model of socioeconomic development in today's advanced economies. As I put it more than a decade ago, "Contemporary economics does not yet have the tools required for a comprehensive and evolutionary theory of economic behavior that would take appropriate account of the main lines of institutional change" (p. 146).[19] Schumpeter (1947) came closer in this century to pro-

[18] Work has been done in some of these areas, but this work has largely lacked a time dimension.

[19] The next several paragraphs borrow heavily from pp. 138–42 of this paper.

viding such a theory than has any other economist in the more-or-less orthodox tradition. Outside that tradition, we can turn to the Marxian literature.

Here, of course, I am harking back to a major plea of the early institutionalists. Veblen asked long ago, in the title of one of his papers, "Why Is Economics Not an Evolutionary Science?," and Wesley Mitchell urged the need for a comprehensive theory of economic behavior that takes the cumulative change of institutions as its chief concern. The institutionalists themselves did not develop such a comprehensive and evolutionary theory. While the classical writers did have the elements of a dynamic system—what Baumol once referred to as the "magnificent dynamics" of the classical school—they theorized on the assumption of a fixed set of social and economic institutions. As Schumpeter put it, "the classics reasoned in terms of a particular historical situation which they uncritically idealized and from which they uncritically generalized" (1947, p. 75).

As for contemporary economists in the neoclassical tradition, they, like their predecessors, seem to be afraid to ask the really big questions about the economic aspects of society—questions which, because they are big, must be concerned with the changing institutional fabric. Some exception to this generalization should be made for the considerable effort that has gone into the study of the underdeveloped parts of the world. Here economists have not been able to ignore the interaction of the institutional environment and economic behavior, and increasing attention has come to be paid to the conditions necessary for one or another kind of change in that environment. I might add here that I continue to be impressed by the fact that in general economists in the advanced countries seem to be prepared to be more institutional in dealing with other parts of the world than they are in studying the particular societies in which they live and do most of their work.[20]

Might the following conceptual framework provide a basis for a more systematic study of the dynamic interaction of economic behavior and the institutional framework? At the most basic level, a society is composed of individual human beings. These individuals are members of households. The larger number of them sell the factor services they control to producing units ("firms" for short); and those who sell labor services must physically participate in the production process. A flow of newly produced goods and services results. The distribution of these goods and services among potential claimants depends on much more than the operation of "impersonal market forces." It reflects a complex of institutional arrangements, which include, among other things, the distribution of power among different groups to influence particular commodity and factor markets, both directly and through government, how the ownership of wealth is distributed and for whose benefit it is used, the tax structure and network of government regulations that emerge from the political process, and the total and distribution of net claims by the rest of the world against domestic output.

Individuals not only are members of households and suppliers of input services to producing units (which may be governmental as well as private), but they are also part of a political process which, while

[20] Lance Davis and Douglass North have recently set forth a tentative and limited "theory of institutional change" and have applied the resulting model to show how a range of economic needs and opportunities have led to new institutional arrangements in the American economy. They are frank in setting forth some of the limitations of their model, and there is a good deal in their approach with which I should quarrel; but their work is clearly a step in the right direction. Incidentally, this particular approach to economic history tends to run counter to the recent quantitative emphasis in that field. Clearly we need to have both types of historical work.

partly local and regional, culminates in the powers of a national government.

Thus we begin with the households and producing units of conventional economic theory but immediately add government as a third basic unit. Households, firms, and government interact within a set of evolving economic institutions. These economic institutions include a hierarchy of markets for current output and for the services of labor, capital, and natural resources, and an array of supporting institutions—from commercial banks to labor mediators to government agencies—that also make a contribution to total output. Households and firms interact not only in response to the standard market stimuli but also by organizing pressure groups to influence both government and particular markets.

These pressure groups—not only labor unions and trade associations but a wide variety of other formal and informal groups—operate within and are conditioned by an evolving set of political and legal institutions that lay down the ground rules as to the way households, firms, and government interact with each other. A major aim of the pressure groups is to influence government and markets not only directly but also by altering legal and political institutional arrangements. And as a result of these pressures, the conditions under which households and firms carry out their economic functions change with the passage of time—as do the ways in which households and firms, through a variety of forms of organization, seek to change these conditions still further in favor of their particular interests.[21]

[21] My emphasis on pressure groups is, of course, much more pluralistic than the Marxist analysis in terms of classes and class conflict. One can synthesize the two approaches by sorting the pressure groups into (two or more) broad classes. But I think that we have to start with the pressure groups if we want to study the functioning of markets and the dynamic interaction between the private sector and government.

We may speak of households, firms, and government as the primary economic agents which carry on their activities within the framework of a set of evolving economic institutions. But these agents and economic institutions also interact with an external environment which can be classified in a variety of ways. One simple classification might be: (1) the framework of legal and political institutions, to which I have already referred; (2) the complex of social institutions that make up what may loosely be referred to as the social environment; (3) the evolving body of scientific and technical knowledge (and the institutions through which such knowledge is developed and transmitted); (4) the physical environment; and (5) the complex of political and economic arrangements that tie a nation to the rest of the world.

Against this background, let us now come back to my final question about the evolving institutional environment, which bluntly put was: How did we get to where we are, and where are we going? Or, if you will, what is the future of capitalism and of the kind of market economy to which we are accustomed and which is changing before our eyes? Lacking a dynamic, politico-economic, and institutionally oriented model, the neoclassical economist averts his eyes—or possibly, and reluctantly, refers the questioner to the still growing Marxist literature. Has not the time come for "orthodox" economists—both defenders and critics of neoclassical theory—to repair this glaring deficiency? Let us borrow what seems appropriate from Marx and his followers as well as from others, although what Marx had to say fitted nineteenth century England much better than it does late twentieth century Western Europe or the United States. But at least let us try to construct a model of the sort I suggest that will have something to say about the evolution and future of the kind of economy and society in which we live.

Sketchy as it is, and it is certainly sketchy, I think my suggested conceptual framework or something similar offers a possible starting point. And there are many intermediate questions to be raised along the way. We have been witnessing a significant extension of government control of the market mechanism in all of the advanced economies, more so in some than in others. This intervention ranges from conventional forms of regulation of particular industries, to sporadic attempts to impose one or another kind of incomes policy, to large-scale programs to redistribute incomes, to widening experiments in worker participation, to outright nationalization of particular firms or industries. What combinations of pressures have caused this extension of government intervention; what forces will extend it further; what forms will such intervention take; and what are likely to be the effects on the allocation of resources, the distribution of income and wealth, and the rates of inflation and of growth in total output—not to mention the possible effects on the various dimensions of the institutional environment, including the institution of private property?

There are many other elements, besides the few mentioned here, that would have to be incorporated into a full-fledged, institutionally oriented theory of economic development for the advanced economies of the nonsocialist world. And we need also to fit in the underdeveloped countries. Here clearly we have to be *political* economists, and not just economists in the neoclassical sense. We are currently witnessing powerful political forces at work aimed at improving the terms of trade of the third world with the advanced countries. What will determine the eventual outcome? Mention of the third world raises a host of other issues, political as well as economic. These include the seemingly inexorable advance of socialism in many of these countries;[22] the effect of autocratic forms of government on the pace of economic growth, the allocation of resources, and the distribution of income; and the ability of these countries to deal with the population growth resulting from high birth rates and declining death rates.

And finally, to repeat, there is for economists the basic question to which not only Marx and his followers but also Schumpeter addressed themselves: What is the future of capitalism in the advanced economies, given the growing size and bureaucratization of business firms, the increasing strength of organized pressure groups, and the momentum from the increasing government intervention that has already occurred? It seems to me that capitalism as we know it in this country or even in Western Europe has little future in the third world. What is its future in the advanced economies?

## IV

And so, on this somber note, I end. I have scolded economists for what I think are the sins that too many of them commit, and I have tried to point the way to at least partial redemption. This road to salvation will not be an easy one for those who have been seduced by the siren of mathematical elegance or those who all too often seek to test unrealistic models without much regard for the quality or relevance of the data they feed into their equations. But let us all continue to worship at the altar of science. I ask only that our credo be: "relevance with as much rigor as possible," and not "rigor regardless of relevance." And let us not be afraid to ask—and to try to answer—the really big questions.

[22] Which leads at least this observer to wonder how compatible are democracy and capitalism in an underdeveloped country whose aspirations are rising much faster than its means to satisfy them.

## REFERENCES


K. J. Arrow, "The Firm in General Equilibrium Theory," in R. Marris and A. Wood, eds., *The Corporate Economy: Growth, Competition and Innovative Potential,* Cambridge, Mass. 1971, 68–110.

———, "Samuelson Collected," *J. Polit. Econ.,* Oct. 1967, *75,* 730–37.

W. J. Baumol, "Monopolistic Competition and Welfare Economics," *Amer. Econ. Rev. Proc.,* May 1964, *54,* 44–52.

G. S. Becker, *Human Capital,* 2d ed., New York 1975.

A. S. Blinder, *Toward an Economic Theory of Income Distribution,* Cambridge, Mass. 1974.

M. Bronfenbrenner, *Income Distribution Theory,* Chicago 1971.

H. Dalton, *Some Aspects of the Inequality of Incomes in Modern Communities,* London 1920.

L. E. Davis and D. C. North, *Institutional Change and American Economic Growth,* London 1971.

C. E. Ferguson and E. J. Nell, "Two Books on the Theory of Income Distribution: A Review Article," *J. Econ. Lit.,* June 1972, *10,* 437–53.

E. Furubotn and S. Pejovich, "Property Rights and Economic Theory: A Survey of Recent Literature," *J. Econ. Lit.,* Dec. 1972, *10,* 1137–62.

J. K. Galbraith, *Economics and the Public Purpose,* Boston 1973.

———, *The New Industrial State,* Boston 1967.

R. A. Gordon, "Institutional Elements in Contemporary Economics," in C. E. Ayres et al., *Institutional Economics: Veblen, Commons, and Mitchell Reconsidered,* Berkeley 1963, 123–47.

R. J. Gordon, "Recent Developments in the Theory of Inflation and Unemployment," presented at Conference of the International Economic Assn., Saltsjöbaden, Sweden, Aug. 1975; to be published in *J. Monetary Econ.,* Apr. 1976.

B. Hansen, *A Survey of General Equilibrium Systems,* New York 1970.

W. W. Heller, "What's Right with Economics?," *Amer. Econ. Rev.,* Mar. 1975, *65,* 1–26.

P. L. Joskow, "Firm Decision-making Processes and Oligopoly Theory," *Amer. Econ. Rev. Proc.,* May 1975, *65,* 270–79.

N. Kaldor, "The Irrelevance of Equilibrium Economics," *Econ. J.,* Dec. 1972, *82,* 1237–55.

T. C. Koopmans, *Three Essays on the State of Economic Science,* New York 1957.

J. Kornai, *Anti-equilibrium. On Economic Systems Theory and the Tasks of Research,* Amsterdam 1971.

W. Leontief, *Essays in Economics: Theories and Theorizing,* New York 1966.

———, "Theoretical Assumptions and Nonobserved Facts," *Amer. Econ. Rev.,* Mar. 1971, *61,* 1–7.

R. Marris, *The Economic Theory of 'Managerial' Capitalism,* London 1964.

J. E. Meade, *Efficiency, Equality and the Ownership of Property,* London 1964.

J. Mincer, "The Distribution of Labor Incomes: A Survey with Special Reference to the Human Capital Approach," *J. Econ Lit.,* Mar. 1970, *8,* 1–26.

W. C. Mitchell, "The Prospects of Economics," in R. G. Tugwell, ed., *The Trend of Economics,* New York 1924, 3–34.

G. Myrdal, *Against the Stream: Critical Essays on Economics,* New York 1973.

A. M. Okun, *Equality and Efficiency: The Big Tradeoff,* Washington 1975.

———, "Upward Mobility in a High-pressure Economy," *Brookings Papers,* Washington 1973, *1,* 207–52.

J. Pen, *Income Distribution: Facts, Theories, and Policies,* New York 1971.

E. S. Phelps et al., *Inflation Policy and Unemployment Theory,* New York 1972.

———, "The New Microeconomics in Inflation and Employment Theory," *Amer. Econ. Rev. Proc.,* May 1969, *59,* 147–60.

E. H. Phelps Brown, "The Underdevelopment of Economics," *Econ. J.,* Mar. 1972, *82,* 1–10.

A. M. Rivlin, "Income Distribution—Can Economists Help?," *Amer. Econ. Rev.*

*Proc.*, May 1975, *65*, 1–15.
J. A. Schumpeter, *Capitalism, Socialism, and Democracy*, 2d ed., New York 1947.
———, *History of Economic Analysis*, New York 1954.
M. Shubik, "A Curmudgeon's Guide to Microeconomics," *J. Econ. Lit.*, June 1970, *8*, 405–34.
———, "Oligopoly Theory, Communication, and Information," *Amer. Econ. Rev. Proc.*, May 1975, *65*, 280–83.
G. J. Stigler, *Essays in the History of Economics*, Chicago 1965.
T. Veblen, "Why Is Economics Not an Evolutionary Science?", *Quart. J. Econ.*, July 1898, *12*, 373–97.
M. L. Wachter, "Primary and Secondary Labor Markets: A Critique of the Dual Approach," *Brookings Papers*, Washington 1974, *3*, 637–80.
B. Ward, *What's Wrong with Economics?*, New York 1972.
P. Wiles, "Cost Inflation and the State of Economic Theory," *Econ. J.*, June 1973, *83*, 377–98.
G. D. N. Worswick, "Is Progress in Economic Science Possible?," *Econ. J.*, Mar. 1972, *82*, 73–86.

# Uncertainty, Permanent Demand, and Investment Behavior

*By* ELEANOR M. BIRCH AND CALVIN D. SIEBERT*

All investment theories must confront two interrelated questions: how the firm decides upon its optimal or desired capital stock; and how the capital stock is adjusted whenever it differs from the optimal or desired level. Most of the current investment studies assume some static or dynamic certainty model which yields an optimal or desired capital stock.[1] In the real world, it is observed that changes in capital stock are more hesitant, more sluggish than changes in this optimal capital stock would suggest. These studies explain this discrepancy in terms of the adjustment process. Entrepreneurs cannot change capital stock as quickly as they wish because of technologically induced lags, financial constraints, specific costs of adjustment, etc.[2] Such approaches have led to distributed lag formulations.

Because of the durable nature of physical assets and the fact that their accumulation and decumulation involve time and require cost outlays, it is axiomatic that the firm's investment in physical assets depends on the future value of economic variables. However, the future is unknown and any views the entrepreneur has on the future must be uncertain. Unfortunately, the implications of such uncertainty have not been explicitly incorporated into studies of investment behavior.

One purpose of this paper is to incorporate the impact of uncertainty on the firm's investment behavior. When uncertainty is given a realistic weight, it affects the firm's equilibrium conditions and makes it more hesitant, more sluggish with respect to changes in its capital stock than it would otherwise be. The more variable expected demand is, the more hesitant the response of optimal capital stock will be.

The incorporation of uncertainty has additional implications for statistical models that purport to explain or predict investment. If a firm operates within a framework of uncertainty, it appropriately views its demand as consisting of permanent and transitory components, and the transitory portion may be impossible to forecast. Especially in its investment decisions, which are of a long-run nature, the firm may plausibly regard the transitory part of demand as a random error and thus focus on the stable, predictable part, its anticipated permanent demand. But if this is so, then investment estimating equations that use actual demand, actual sales, actual prices, etc., as independent variables are subject to the biases inherent in errors-in-variables formulations.[3] The nature of these biases should help to explain the ob-

* Associate professor, department of business administration, and professor, department of economics, University of Iowa, respectively. An earlier draft of this paper was presented at the meetings of the Econometric Society, New York, December 1973. Financial support by the Ford Foundation and the University of Iowa is gratefully acknowledged.

[1] For a survey of recent investment literature, see Dale Jorgenson. For investment studies prior to 1960, see Robert Eisner and Robert Strotz.

[2] For technologically induced lags, see Eisner (1960); for financial constraints, see John Meyer and Edwin Kuh, and James Duesenberry; and for costs of adjustment, see Eisner and Strotz, Arthur Treadway, Hyman Joseph, and Robert Lucas.

[3] The classical discussion of the errors-in-variables problem is in Maurice Kendall and Alan Stuart. For further developments, see Dennis Lindley. For specific examples of the errors-in-variables model applied to economic problems, see Milton Friedman; Gregory Chow, especially Appendix I, pp. 94–98; Tony Lancaster; and Merton Miller and Franco Modigliani.

served underestimation of the elasticity of capital stock with respect to output in distributed lag accelerator regressions.

That firms might, in fact, behave this way was suggested by Eisner (1967), who pointed out the relevance for investment studies of Friedman's permanent income hypothesis and suggested some of the implications of the errors-in-variables formulation for empirical investment studies. Our paper builds on Eisner's suggestions and makes permanent demand the relevant concept for entrepreneurial decisions on investment.

The above topics are approached in the following framework. In the first section the implications for investment behavior of the introduction of costs-of-adjustment and uncertainty are developed. In Section II the problem of changing expectations in an uncertain context is discussed. This leads to the errors-in-variables model as the appropriate statistical approach to the estimation of investment functions. In Section III the statistical model is applied to the accelerator model of investment behavior. The exploratory empirical application confirms the implications of the permanent demand investment model, viz., the estimate of elasticity of capital stock with respect to sales increases significantly; and the length of the estimated average lag decreases when the errors-in-variables estimating approach is employed.

## I. Costs of Adjustment, Uncertainty, and Investment Behavior

In a world of certainty with a perfect capital market, exponentially evaporating capital stock, and costless adjustment of capital stock, the optimal investment behavior of the firm is the relatively simple one of holding for each instant of time that stock of capital for which the marginal revenue product of capital equals the cost of capital.[4] This would imply that if capital stock were pictured as a function of time, it would present a jagged appearance, fluctuating with every vagary of demand. If, more realistically, we assume some costs of adjustment, the capital stock series would become smoother. Even an entrepreneur with perfect foresight would not keep his capital stock perfectly adjusted to optimal output in each period. He would hesitate to increase capital stock even when demand increased, unless the increase were sustained over a long enough period to justify the cost of increasing, then decreasing, capital stock.

This inertia or rigidity in capital stock is reinforced when uncertainty is introduced into the problem.[5] In reality, the entrepreneur does not know the level of demand in any future time period; he must estimate it. When an entrepreneur plans his investment expenditures, he is choosing future levels of capital stock, which depend on his expectations about future demand. But these expectations are formed under uncertainty; thus, there is an additional opportunity cost associated with the investment decision, viz., the cost of making an incorrect estimate about future demand. Moreover, since the entrepreneur's perception of demand is hazy under uncertainty, a shift in demand due, say, to a change in consumer tastes, is difficult to distinguish from a random fluctuation. Thus, the entrepreneur's response to such a change is hesitant, and especially so with respect to the investment decision, which depends on longer range (i.e., more uncertain) rather than shorter range forecasts.

In such an environment, the investment decision can be viewed as follows. The entrepreneur makes a long-range forecast of demand (for example, one year) which implies a series of short-range forecasts (for example, monthly during the year).

[4] See Kenneth Arrow for a demonstration of these results.

[5] For further development of this notion, see the paper by Guy Stevens. Also, the issue is treated in a longer version of this paper available from the authors on request.

The entrepreneur decides on his optimal capital stock on the basis of the long-range forecast, and the difference between this and his current capital stock determines his investment. He then proceeds to produce from month to month on the basis of his short-run equilibrium conditions and his original long-range forecast, as long as the forecast is validated by events. The entrepreneur's problem then becomes how to decide when a forecast has become invalid. One way to describe how the entrepreneur behaves under these conditions is to picture him as adopting a hypothesis-testing approach. That is, the entrepreneur continues to accept the null hypothesis that demand has not permanently changed unless events convince him otherwise.

If the entrepreneur accepts the null hypothesis, he continues to leave his capital stock unchanged; but if the entrepreneur rejects it, he makes a new long-range forecast of permanent demand, chooses a new optimal capital stock on the basis of this forecast, and invests (disinvests) accordingly. This behavior on the part of the firm suggests that minor fluctuations in demand (those consistent with the long-range forecast) are met by moving along the short-run cost curve while significant changes in demand cause a long-run response. That is to say, the entrepreneur balances the opportunity loss he risks by making a wrong decision on capital stock against the opportunity loss he makes by operating on his higher short-run cost curve rather than on his lower long-run curve. As long as the former exceeds the latter, the entrepreneur leaves his capital stock unchanged. When the two reverse positions, so that the latter exceeds the former, the entrepreneur will invest (disinvest) so as to change capital stock.

Thus the entrepreneur in the face of uncertainty in conjunction with positive costs-of-adjustment will be quite sluggish in making changes in capital stock. A decision to change capital stock will be made only after the entrepreneur decides demand has changed permanently. Such a decision may take several periods of observation. Even after a decision has been made that permanent demand has changed, costs-of-adjustment may deter an immediate movement to the entrepreneur's optimal capital stock.

## II. Expectations, Permanent Demand, and Errors in Variables

Looking at the decision problem for the entrepreneur faced with uncertainty and high costs of adjustment as one in which the entrepreneur makes a new long-run forecast when he rejects the hypothesis that permanent demand has not changed, one is faced with the question of how to capture the changes in the entrepreneur's expectations for purposes of estimation. One approach to this problem is the adaptive expectations formula of Phillip Cagan and Marc Nerlove which has become quite popular. As Edwin Mills shows, however, such an autoregressive model implies biased expectations. Mills demonstrates that any predictive information in an autoregressive scheme will be better utilized if it is built into the vector of explanatory variables. Thus a regression equation using the expectational variable as the dependent variable and any other explanatory variables we think relevant as the independent variables (including, if desired, past values of the expectational variable itself) should provide a better estimate. If, for example, we assume a fixed-price model, then anticipated permanent demand for some future period consists of anticipated permanent sales. If we make the assumption that on the average entrepreneurs guess correctly about their future sales, then we can regress actual sales on a set of explanatory variables, thus obtaining regression coefficients with which to form a predictive equation for anticipated permanent sales. If some such scheme is not followed, but instead actual

sales are used as an independent variable in an investment equation instead of anticipated permanent sales, then the statistical problem of errors-in-variables is likely to be encountered.[6]

The relevance of the errors-in-variables analysis for investment studies is obvious. Take, for example, the accelerator model of investment. In its most general form it focuses on a generally accepted economic doctrine, viz., that the demand for capital is a derived demand.[7] But it is not necessary to link investment with historical changes in sales to preserve this truth. It is really anticipated long-run optimal sales, not historical output, that is functionally related to optimal capital stock. Investment then depends on changes in anticipated long-run optimal sales, which do not necessarily occur *pari passu* with actual sales changes especially where imperfect knowledge is the rule. Optimal capital stock will tend to be more sluggish under uncertainty than actual output. This suggests that investment depends on changes in permanent demand (or, in a fixed-price model, permanent sales) rather than in actual demand (sales) under the assumption of a fixed system of cost curves.

Now if this view of investment is valid, it suggests that traditional formulations of the investment equation based on the accelerator theory lead to inconsistent estimates of the parameters of that function, in just the same way that consumption functions based on measured income yield inconsistent estimates of the marginal propensity to consume. Investment literature abounds with accelerator-type regressions where the elasticity of capital stock with respect to output is considerably lower than the postulated figure of unity. Many possible explanations are proffered: (a) the firm was not operating at full capacity so the full accelerator was not operative; (b) other factors, for example, the firm's liquidity, must also be taken into account; and (c) the firm is never in equilibrium because of various costs of adjustment. Reason (c) led to the various distributed lag formulations. Now many or even all of these explanations could be true in some particular case, but the existence of the classic errors-in-variables problem may well explain part of the discrepancy.

Consider an accelerator-type investment equation in which investment in year t is regarded as a linear function of the change in sales in the previous year as well as of other explanatory variables such as a capacity indicator, profits, etc., the latter variables added to account for times when the strong accelerator is inoperative:

$$I_t = b_0 + b_1(S_t - S_{t-1}) + \sum_{i=2}^{m} b_i X_{i,t-1} + v_t \tag{1}$$

To the extent that the actual sales change in (1) is a proxy for the expected sales change, the coefficient $b_1$ will be biased downward. The "true" relation would be:

$$I_t = \beta_0 + \beta_1(S_t^* - S_{t-1}^*) + \sum_{i=2}^{m} \beta_i X_{i,t-1} + u_t \tag{2}$$

where $E(u_t) = E(u_{t-r}) = 0$, for all $r$;

$$cov\,(u_{t-i}, u_{t-j}) = \begin{cases} \sigma_u^2, & i = j \\ 0 & \text{elsewhere} \end{cases}$$

**But**

$$\begin{aligned} S_t &= S_t^* + e_t \\ S_{t-1} &= S_{t-1}^* + e_{t-1} \end{aligned} \tag{3}$$

[6] The earliest econometric efforts were usually errors-in-variables models, but for the past generation, errors-in-equation models have superseded them. Friedman's consumption function based on permanent income rejuvenated the errors-in-variables approach.

[7] The accelerator model is employed since the statistical approach developed here can be applied in a more straightforward manner than in other models of investment behavior. Also, Eisner (1967) suggested the errors-in-variables approach for the accelerator model.

where $E(e_t)=E(e_{t-r})=0$, for all $r$;

$$cov\,(e_{t-i}, e_{t-j}) = \begin{cases}\sigma_e^2 & i=j\\ 0 & \text{elsewhere}\end{cases}$$

Therefore, $S_t^*-S_{t-1}^*=S_t-S_{t-1}-(e_t-e_{t-1})$ where $E(e_t-e_{t-1})=0$ and $var\ (e_t-e_{t-1})=2\sigma_e^2$. If (1) is estimated when (2) is the true model, then (1) becomes indistinguishable from:

$$(4)\quad I_t = \beta_0 + \beta_1(S_t - S_{t-1}) + \sum_{i=2}^{m} \beta_i X_{i,t-1} + u_t - \beta_1(e_t - e_{t-1})$$

or

$$I_t = \beta_0 + \beta_1(S_t - S_{t-1}) + \sum_{i=2}^{m} \beta_i X_{i,t-1} + v_t$$

where $v_t=u_t-\beta_1(e_t-e_{t-1})$. Then the covariance between the error term and the independent variable (under the usual assumption that $u$ and $e$ are independent) is

$$(5)\quad E\{v_t[(S_t - S_{t-1}) - E(S_t - S_{t-1})]\} = -2\beta_1\sigma_e^2$$

which will be zero only when there is no measurement error in the independent variable or in the trivial case where $\beta_1=0$. In general, this covariance will be nonzero and therefore (1) will be subject to the equivalent of the familiar simultaneous equations bias. In the limit, for large samples, the coefficient of actual sales change $b_1$ will be biased downward as an estimator of $\beta_1$. The other regression coefficients are also affected, being biased upward (downward) if their variable is positively (negatively) related to permanent sales changes.[8]

From the above discussion, it would seem that the model dealt with here is actually a system of simultaneous equations:

$$(6a)\quad I_t = \beta_0 + \beta_1(S_t^* - S_{t-1}^*) + \sum_{i=2}^{m} \beta_i X_{i,t-1} + u_t$$

$$(6b)\quad S_t = S_t^* + e_t$$
$$S_{t-1} = S_{t-1}^* + e_{t-1}$$

$$(6c)\quad S_t^* = \gamma_1 + \sum_{i=2}^{m} \gamma_i X_{i,t-1} + \omega_t$$
$$S_{t-1}^* = \gamma_1 + \sum_{i=2}^{m} \gamma_i X_{i,t-2} + \omega_{t-1}$$

where $I_t$= net investment in year t; $S_t^*$= true but unobservable permanent sales in year t; $S_t$= actual sales in year t; $X_i$, $i=2, \ldots, m$ stand for all other relevant variables, some $\beta_i$ and $\gamma_i$ may be zero; and with the error terms assumed independent of each other and with means zero and constant variances of $\sigma_u^2$, $\sigma_e^2$, and $\sigma_\omega^2$, respectively. Equations (6) thus constitute a system in which both $I$ and $S$ are treated as endogenous; only the $X_i$ are truly exogenous variables. When (6a) is estimated with actual sales used in place of permanent ones, actual sales are erroneously treated as exogenous variables, leading to the familiar simultaneous equations bias.

One way to reduce the bias in $\beta_1$ is to follow the instrumental variable approach suggested by equations (6). First the endogenous variable $S_t$ can be regressed on all the instrumental variables $X_i$ in (6c). From these estimates a new variable, $\hat{S}_t$, is formed, defined as:

$$(7)\quad \hat{S}_t = g_1 + \sum_{i=2}^{m} g_i X_{i,t-1}$$

This variable provides an estimate of $S_t^*$ from which much of the measurement error $e_t$ will have been removed. If $\hat{S}_t$ is then used in the second stage as the sales variable in place of $S_t^*$, with $\hat{S}_{t-1}$ for $S_{t-1}^*$, the resulting estimates of the $\beta_i$ should be more accurate estimates of the $\beta_i$ in the basic

[8] See Chow, pp. 94–98, for detailed proofs.

structural equation (6a). The task then is to estimate (6a), first using $(S_t - S_{t-1})$ as the main independent variable, then note the difference when it is estimated with $(\hat{S}_t - \hat{S}_{t-1})$ as the main independent variable, in a two-stage procedure. The result should be that the coefficient of the sales-change variable should increase in the second stage.

The two-stage procedure can also shed some light on an old controversy in the investment literature, viz., the role of profits in investment. Jan Tinbergen held that the prospect of future profits induces investment and that current profits are an indicator of future profits. Thus profits enter the investment picture on the demand side; i.e., they directly influence the firm's desire to expand. There is another school of thought about profits, however, associated with Meyer and Kuh and with Duesenberry, among others, which maintains that profits affect the supply of funds that finance investment. This notion depends on the assumption that external capital rationing exists or that firms, for this or whatever reason, prefer to finance their expansion out of retained earnings.

The usual course taken in investment studies is simply to add a profits variable (or several) with unspecified coefficient to the structural equation and let the sample determine its value. The trouble with this approach, however, is that if applied in this simplistic fashion, the resulting estimate of the profits coefficient (in the liquidity, or supply-side sense) will, because of the errors-in-variables problem, inevitably be upward biased and the sales-change coefficient correspondingly downward biased. To the extent that Tinbergen's view is valid, the profits coefficient in the regression equation will reflect substantial informational content about expected earnings (which, in this model, is reflected by $S_t^*$, expected sales) along with the supply effect, if any, of profits per se on investment. This could help to explain some empirical results obtained in many studies with respect to profits. When profits have been included as an explanatory variable along with actual sales changes, their coefficients have usually been positive and sometimes significantly so. But this is consistent with a true value of zero for the profits coefficient (in the liquidity sense) if profits are positively correlated with expected sales, as Tinbergen would maintain. This confounding of profits' supply and demand effects can be reduced by a two-stage procedure and some, at least, of the demand-side content of profits can be salvaged. To the extent that profits contribute anything to equation (6c), their coefficient should be lower when $(\hat{S}_t - \hat{S}_{t-1})$ is used in (6a) than it is when $(S_t - S_{t-1})$ is used. The profits coefficient in (6c) should provide some idea of their demand-side influence, and their coefficient in (6a)—when $(\hat{S}_t - \hat{S}_{t-1})$ is used—should indicate their liquidity or supply-side influence.

Another implication of this approach is that if the method succeeds in estimating anticipated permanent demand, the average lag that can be derived from the estimated investment equations should be shorter when anticipated permanent sales changes are used than when actual sales changes are entered into the regression.

## III. Empirical Results

To test the model developed above, a pilot study of a group of eight large *U.S.* industrial firms was undertaken. These firms were chosen mainly because 1934–63 data were available for them in a particularly convenient form, as a result of earlier studies by Jorgenson and Siebert.[9]

[9] The data sources and measurement methods are the same as those of the Jorgenson and Siebert studies. The firms used in this study were: 1) American Can Co., 2) General Electric Co., 3) General Motors Corp., 4) International Business Machines, 5) R. J. Reynolds Tobacco Co., 6) Standard Oil Co., 7) Swift and Co., and 8) U.S. Steel Corp. Details on data measurement are

These figures were then extended through 1966 except in a few instances where the company's method of accounting had changed and the data would no longer be comparable.

The variables used in the regressions are defined as follows:

$S_{t-1}$ = the firm's sales in year t−1.

$\hat{S}_t$ = the firm's predicted permanent sales for year t estimated on the basis of knowledge through year t−1.

$P_{t-1}$ = the firm's profits before taxes in year t−1.

$S^I_{t-1}$ = the industry's sales in year t−1.

$MVF_{t-1}$ = the market value of the firm at the end of year t−1. This is the sum of the firm's debt and the market value of its equity. Yehuda Grunfeld used this as an indicator of expected future profits.

$I_t$ = the firm's investment in fixed plant and equipment during year t.

$D_t$ = estimated replacement, i.e., "true" depreciation for the firm in year t.[10]

The first-stage set of regressions consisted of four equations estimating permanent sales. These estimates were assumed to be made before the fact, i.e., under conditions of uncertainty. The instrumental variables used were the firm's own sales, its profits, its market value, and the sales of its industry. In the first equation these independent variables were expressed as three-year averages in order to remove transitory effects. For the firm's own sales, this choice seems reasonable, but for the other variables it is possible that changes in their levels or even their growth rates might be more meaningful. For example, a firm might be more inclined to believe its sales would increase if its profits were increasing rather than merely high, and it might be more influenced by a high growth rate of its industry's sales rather than by its current high level. Accordingly, the second equation used these instrumental variables expressed as a change in their levels from year t−2 to year t−1 and the third used them expressed as growth rates or percentage changes from year t−2 to t−1. Thus, the first three estimating equations were as follows:

$$(8)\quad (\hat{S}_t) = g_0 + g_1[1/3(S_{t-1}+S_{t-2}+S_{t-3})] + g_2[1/3(P_{t-1}+P_{t-2}+P_{t-3})] + g_3[1/3(S^I_{t-1}+S^I_{t-2}+S^I_{t-3})] + g_4 MVF_{t-1}$$

$$(9)\quad (\hat{S}_t) = g_0 + g_1[1/3(S_{t-1}+S_{t-2}+S_{t-3})] + g_2(S_{t-1}-S_{t-2}) + g_3(P_{t-1}-P_{t-2}) + g_4(S^I_{t-1}-S^I_{t-2}) + g_5(MVF_{t-1}-MVF_{t-2})$$

$$(10)\quad (\hat{S}_t) = g_0 + g_1[1/3(S_{t-1}+S_{t-2}+S_{t-3})] + g_2\left(\frac{S_{t-1}-S_{t-2}}{S_{t-2}}\right) + g_3\left(\frac{P_{t-1}-P_{t-2}}{P_{t-2}}\right) + g_4\left(\frac{S^I_{t-1}-S^I_{t-2}}{S^I_{t-2}}\right) + g_5\left(\frac{MVF_{t-1}-MVF_{t-2}}{MVF_{t-2}}\right)$$

The results of these first three equations were then used to decide upon a fourth and

contained in a longer version of this paper available from the authors on request.

[10] Given investment at constant prices and two benchmark capital stock figures, all other capital stock and replacement data were computed from an explicit model for replacement, assuming that replacement was a constant fraction of capital stock at the start of each year. This yielded a difference equation that could be solved for an estimate of the "true" depreciation rate. This in turn was used to compute a "true" depreciation series.

final version. The fourth equation used as independent variables those that were found to be significant in the first three formulations. Thus, permanent sales in year t were estimated as a function of the firm's average sales in the preceding three years, a variable whose coefficient differed significantly from zero in all three earlier versions; the firm's market value at the beginning of year t, significant in the first equation; the change in the firm's profit level during the preceding year, significant in the second equation; the change in the firm's market value during the preceding year, significant in the second equation; the growth rate of the industry's sales during the preceding year, significant in the third equation; and lastly, the firm's profits in the preceding year. The profits variable was added in order to separate the demand-side, informational contribution of the profits variable from the liquidity or supply-side contribution it might make to the subsequent second-stage investment equations. Thus, the fourth equation was of the form:

$$(11)\quad (\hat{S}_t) = g_0 + g_1[1/3(S_{t-1} + S_{t-2} + S_{t-3})] + g_2 MVF_{t-1} + g_3(P_{t-1} - P_{t-2}) + g_4(MVF_{t-1} - MVF_{t-2}) + g_5\left(\frac{S^I_{t-1} - S^I_{t-2}}{S^I_{t-2}}\right) + g_6 P_{t-1}$$

The regression program used for the above four equations had a stepwise feature which permitted each variable to enter the regression singly, in order of importance. For example, if one independent variable was already in the equation, the next variable to be added to the equation would be the one with the highest $F$-level for entrance. Because of multicollinearity among the explanatory variables, the addition of a third or fourth variable often added virtually nothing in the way of fresh explanation; its main effect was to render insignificant the coefficients of variables previously added. To rule out these overly collinear effects, the rule adopted was to accept as final the equation where any omitted variable, if it were to enter, would render insignificant the coefficient of any variable in the equation, including its own (at a significance level of 95 percent).

The results for all four equations predicting permanent sales are shown in Table 1. The independent variables shown in the table are considerably fewer than those shown above because many of those originally considered turned out to be insignificant; only those whose coefficients were significantly different from zero at the 95 percent level are shown in Table 1.

The firm's own sales average was, as expected, most helpful in predicting its permanent sales. Industry sales were less important.[11] But profits entered significantly into two of the equations as did the market value of the firm variable. These results tend to confirm Grunfeld's belief that this variable has expectational content.

Since the fourth equation was the most general in that it allowed the instrumental variables to enter in any form, and since it estimated sales with slightly greater accuracy ($R^2 = .98$) than the other three ($R^2 = .97$), it was selected as the "best" predictor of permanent sales. The fourth equation was then used to generate a series of permanent sales which could be used in the second-stage problem, the investment estimation equation.

The simplest possible version of an accelerator-type investment equation would be one that made gross investment in year

[11] This would appear to contrast with Eisner's (1967, pp. 368, 374) results in which he hypothesizes that firms view variations in their industry's sales as a better predictor of their future prospects than their own past sales experience. However, the firms included in this study are the largest in their industries, so they may dominate the industry influence.

TABLE 1—RESULTS OF FOUR REGRESSIONS PREDICTING PERMANENT SALES IN YEAR t AS A FUNCTION OF SPECIFIED VARIABLES

| | | Regression Coefficients and Standard Errors | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Equation | $R^2$ | Constant Term | $\frac{1}{3}(S_{t-1}+S_{t-2}+S_{t-3})$ | $MVF_t$ | $P_{t-1}-P_{t-2}$ | $MVF_t-MVF_{t-1}$ | $\frac{S^I_{t-1}-S^I_{t-2}}{S^I_{t-2}}$ | $P_{t-1}$ |
| (8) | .97 | 14.00 | 1.019 | .0682 | | | | |
| | | (62.23) | (.0305) | (.0183) | | | | |
| (9) | .97 | 11.55 | 1.089 | | .6257 | .1699 | | |
| | | (59.38) | (.0145) | | (.2048) | (.0377) | | |
| (10) | .97 | −86.00 | 1.124 | | | | 864.4 | |
| | | (69.49) | (.0145) | | | | (282.4) | |
| (11) | .98 | 96.09 | .9113 | .0464 | | | | .7340 |
| | | (64.75) | (.0421) | (.0189) | | | | (.2038) |

t a function of the firm's replacement needs, i.e., its true depreciation, and its current year sales change:

$$I_t = \beta_0 + \beta_1(D_t) + \beta_2(S_t - S_{t-1}) + u_t \tag{12}$$

This is perhaps overly simplified but it does serve to focus attention on the sales-change variable which is of primary interest here. Accordingly, equation (12) was estimated in two forms: first, with actual sales used for $S_t-S_{t-1}$ and then with permanent sales, $\hat{S}_t-\hat{S}_{t-1}$, yielded by equation (11). The results are presented in Table 2.

When the permanent rather than the actual sales change was used as the independent variable, the regression equation explained 73 rather than 47 percent of the variation in investment. The coefficient of the sales change variable rose from .0569 when actual sales were used to .4052 when permanent sales were used; the increase was statistically significant at the 99.5 percent confidence level.

TABLE 2—RESULTS OF THREE PAIRS OF REGRESSIONS ESTIMATING INVESTMENT ACCORDING TO EQUATIONS (12), (17), AND (18), USING ACTUAL SALES AND PREDICTED PERMANENT SALES

| | | Regression Coefficients and Standard Errors | | | | | |
|---|---|---|---|---|---|---|---|
| Equation | $R^2$ | Constant Term | $D_t$ | $S_t-S_{t-1}$ | $I_{t-1}$ | $S_{t-1}-S_{t-2}$ | $P_{t-1}$ |
| (12a) Using actual sales | .47 | 129.3 | .6024 | .0569 | | | |
| | | (21.48) | (.0488) | (.0274) | | | |
| (12b) Using predicted permanent sales | .73 | 68.80 | .5585 | .4052 | | | |
| | | (15.85) | (.0339) | (.0281) | | | |
| (17a) Using actual sales | .93 | 6.865 | .0527 | .0462 | .8909 | .0847 | |
| | | (8.751) | (.0252) | (.0106) | (.0307) | (.0109) | |
| (17b) Using predicted permanent sales | .93 | 7.873 | .1002 | .1629 | .8430 | −.0392 | |
| | | (8.504) | (.0259) | (.0181) | (.0370) | (.0198) | |
| (18a) Using actual sales | .94 | 5.941 | .0241 | .0414 | .6868 | .0467 | .1462 |
| | | (8.072) | (.0237) | (.0098) | (.0440) | (.0119) | (.0242) |
| (18b) Using predicted permanent sales | .94 | 7.320 | .0469 | .0879 | .7223 | −.0540 | .1313 |
| | | (8.159) | (.0277) | (.0245) | (.0452) | (.0288) | (.0304) |

It was suggested above that equation (12) is a rather simplified formulation. Because of costs-of-adjustment it is not optimal for a firm to adjust its capital stock quickly. Such partial adjustment can lead to a flexible accelerator model of investment in which not only the current year's sales changes, but also the preceding years' changes enter the regression.[12] This leads to a distributed lag formulation of the investment equation. Some authors, notably Eisner (1960), have included as many as seven sales-change variables in an investment equation of this type. It is possible to suggest the same process, while economizing on variables, by using some version of a Koyck-type distributed lag.[13] This can be done by assuming that beyond some time period in the past, the influence of prior periods decreases geometrically. For example, net investment in year t, $Q_t$, can be considered to be a linear function of all past sales changes (either actual or permanent). Thus:

$$(13) \quad Q_t = a_0 + a_1\Delta S_t + a_2\Delta S_{t-1} + a_3\Delta S_{t-2} + a_4\Delta S_{t-3} + \ldots$$

Gross investment in year t, $I_t$, would be net investment $Q_t$ plus replacement investment $R_t$, so:

$$(14) \quad I_t = R_t + a_0 + a_1\Delta S_t + a_2\Delta S_{t-1} + a_3\Delta S_{t-2} + a_4\Delta S_{t-3} + \ldots$$

The decreasing geometric series can be introduced at any point. If one wishes, one can estimate any finite number of the $a_i$, $i=0, \ldots, \infty$, assuming that the remaining $a_{i+j}$, $j=1, \ldots, \infty$, are related to $a_i$ as follows:

$$(15) \quad a_{i+j} = h^j a_i, \qquad 0 < h < 1$$

[12] See Eisner and Strotz and Lucas for a demonstration that introducing costs-of-adjustment in the investment function can lead to a flexible accelerator-type investment model.

[13] A Koyck-type distributed lag is sufficient for our purposes since we are not attempting to estimate the best form of the lag function.

For example, if one estimates only the first two sales-change coefficients and permits the geometric decrease to begin beyond that point, one can derive:

$$(16) \quad I_t = (1 - h)a_0 + (R_t - hR_{t-1}) + hI_{t-1} + a_1\Delta S_t + (a_2 - ha_1)\Delta S_{t-1}$$

Because of the presence of $R_t$ and $R_{t-1}$ in this equation, the multicollinearity problem would be severe if it were estimated in this form. Since the replacement (true depreciation) series used here varied very little from one year to the next, $(R_t - hR_{t-1})$ was assumed to be approximately equal to $(1-h)R_t$, so equation (16) above became the following equation to be estimated here:

$$(17) \quad I_t = \beta_0 + \beta_1(D_t) + \beta_2(I_{t-1}) + \beta_3(S_t - S_{t-1}) + \beta_4(S_{t-1} - S_{t-2}) + U_t$$

where $\beta^0 = (1-h)a_0$, $\beta_1 \sim (1-h)$, $\beta_2 = h$, $\beta_3 = a_1$, and $\beta_4 = a_2 - ha_1$.

Equation (17) was then estimated in two forms: first, with actual sales used for $S_t - S_{t-1}$ and $S_{t-1} - S_{t-2}$; and then with permanent sales $\hat{S}_t - \hat{S}_{t-1}$ and $\hat{S}_{t-1} - \hat{S}_{t-2}$, yielded by equation (11). The results are presented in Table 2.

Based on the earlier discussion, we should expect that equation (17b) would explain investment at least as well as—hopefully, better than—equation (17a), and that the coefficient of $S_t - S_{t-1}$ would be larger in (17b) than in (17a). Also, owing to the derivation of equation (17), $\hat{\beta}_4$, the coefficient of $(S_{t-1} - S_{t-2})$, can be used together with $\hat{\beta}_2$ and $\hat{\beta}_3$ to estimate the original distribution lag coefficients $a_1$, $a_2$, and all remaining ones if desired. If this were done, then we should expect that the distributed lag estimated from (17a), the actual sales equation, would be longer than the one estimated from (17b), the permanent sales equation. In other words, the

two-stage procedure should serve to disentangle the purely expectational lag from the costs-of-adjustment one. To the extent that the expectational lag has already been subsumed under the permanent sales variable, there should remain only the costs-of-adjustment lag. So the estimated value of $a_3$ in equation (13) should decrease when permanent sales are substituted for actual sales in the estimation of equation (17). This, together with an increase in $a_1$, would imply the lag has been foreshortened.

As can be seen from Table 2, equation (17b) was at least as good as equation (17a) in explaining investment, since 93 percent of the variation in investment was explained by each. (Actually, equation (17b) was very slightly better with $R^2 = .930$ as compared with $R^2 = .927$ for equation (17a).) The change in the coefficient of $S_t - S_{t-1}$ was also in the predicted direction, increasing substantially when permanent sales were used. The increase was statistically significant at the 99.5 percent confidence level. The estimates of $a_1$, $a_2$, and $a_3$ for the actual sales version were .0462, .1259, and .1122, respectively. For the permanent sales version, they were .1629, .0981, and .0847, so the distributed lag has indeed been shortened by the two-stage procedure.

The investment equations considered thus far have made no allowance for liquidity or supply-side variables. As mentioned earlier, Duesenberry and Meyer and Kuh, among others, have stressed the importance to investment of a high level of profits. The investment equation previously considered could be altered to include a profits variable as follows:

$$
\begin{aligned}
(18)\quad I_t = {} & \beta_0 + \beta_1(D_t) + \beta_2(I_{t-1}) + \beta_3(P_{t-1}) \\
& + \beta_4(S_t - S_{t-1}) + \beta_5(S_{t-1} - S_{t-2}) \\
& + U_t
\end{aligned}
$$

Since equation (18) contained the profits variable[14] that was included in equation (11), the estimate of permanent sales, estimating equation (18) in two ways with actual sales and with permanent sales should serve to sort out the demand-side and supply-side effects of the previous year's profits on investment. Accordingly, this equation was estimated with the results shown in Table 2. The increase in the coefficient of $S_t - S_{t-1}$ when permanent sales were substituted for actual ones was statistically significant at the 95 percent confidence level. The change in the estimated lag structure here was also as predicted. Estimates of $a_1$, $a_2$, and $a_3$ were .0414, .0751, and .0516 for the actual sales equation; they were .0879, .0095, and .0069, respectively, for the permanent sales equation.[15]

The main reason for estimating equation (18) was to observe the behavior of $\hat{\beta}_3$, the coefficient of the profits variable. If

[14] There is some question as to what measure of profits should be included in the second-stage or investment equation regression. Kuh, p. 63, suggested profits after taxes plus depreciation less dividends paid as an appropriate measure for the supply of funds impact on investment. In the results reported here, however, the same profits variable was employed in both stages of estimation—in estimating permanent sales and in the second-stage investment equation.

[15] To rule out the possibility that these results might be spurious owing to the presence of heteroskedasticity, the investment equations were also estimated in ratio form, with each equation divided through by capital stock, an indicator of firm size. The results were not materially changed. The Durbin-Watson statistic was used to check independence; for what it is worth, this gave favorable results (i.e., no evidence of serial correlation) for all equations except (12a) and (12b), which are no doubt misspecified. But the Durbin-Watson test is not strictly applicable to the case where lagged variables are used and it is biased toward a favorable result in that event. Nevertheless, there is reason to believe that autocorrelation in the residuals is not a major problem. The focus of this study was not on the estimation of coefficients for their own sake, but on ascertaining whether the use of different sales variables gives rise to differences in regression coefficients. For each pair of equations compared, any irregularity such as autocorrelation would be expected to occur in both versions. Thus, differences between them which were of primary interest here should not be appreciably affected in any event.

profits have any demand-side influence, i.e., if they affect the estimate of permanent sales, then the coefficient of the profits variable should be lower in equation (18b) than in equation (18a). This is because the demand-side influence of this variable should already be reflected in the permanent sales variable used in (18b) while it is not so reflected in the actual sales variable used in (18a). How much lower the coefficient would become is a moot question, since it depends on the relative strengths of the demand-side and supply-side influences of this variable. As can be seen from Table 2, the profits coefficient did decrease in equation (18b) as predicted but the decrease was not very large. In fact, there is about a one-third chance that the decrease was not statistically significant. From this test, then, it would appear that the profits variable could have both a demand-side and a supply-side influence on investment. Further tests with larger samples should permit better estimates of their magnitudes.

## IV. Concluding Remarks

The results of the empirical analysis were consistent with our expectations: the estimated elasticity of capital stock with respect to sales increased, and the estimated average lag decreased when the instrumental errors-in-variables approach was employed. The empirical results also lend support to the hypothesis that an entrepreneur faced with uncertainty will tend to adjust only when he considers changes to be permanent. Our results suggest that further work with an instrumental errors-in-variables approach for investment behavior may prove fruitful, especially in disentangling supply-side from demand-side effects of profits. Our results are also consistent with those of Eisner (1967). He found that sales expectations of firms have some role in explaining investment over and above that provided by actual current and lagged sales changes.

In addition, the results have implications beyond the confines of investment behavior. Most economic acts are based upon expectations formed about an uncertain future. Under such conditions, economic agents' behavior will tend to depend on anticipated permanent values of relevant decision variables. If that is the case, the two-stage instrumental errors-in-variables approach employed in this paper may be a more appropriate estimation procedure than the conventionally used errors-in-equation model.

## REFERENCES

K. Arrow, "Optimal Capital Policy, The Cost of Capital and Myopic Decision Rules," *Ann. Inst. Statist. Mathemat.*, 1964, *16*, 21–30.

P. Cagan, "The Monetary Dynamics of Hyper-inflation," in M. Friedman, ed., *Studies in the Quantity Theory of Money*, Chicago 1956.

G. Chow, *Demand for Automobiles in the United States*, Amsterdam 1957.

J. Duesenberry, *Business Cycles and Economic Growth*, New York 1958.

R. Eisner, "A Distributed Lag Investment Function," *Econometrica*, Jan. 1960, *28*, 1–29.

———, "A Permanent Income Theory for Investment," *Amer. Econ. Rev.*, June 1967, *57*, 363–90.

——— and R. Strotz, "Determinants of Business Investment," in *Impacts of Monetary Policy: Commission on Money and Credit*, Englewood Cliffs 1963, 59–337.

M. Friedman, *A Theory of the Consumption Function*, Princeton 1957.

Y. Grunfeld, "The Determinants of Corporate Investment," in A. C. Harberger, ed., *The Demand for Durable Goods*, Chicago 1960.

D. Jorgenson, "Econometric Studies of In-

vestment Behavior: A Survey," *J. Econ. Lit.*, Dec. 1971, *9*, 1111–47.

——— and C. Siebert, (1968a) "A Comparison of Alternative Theories of Corporate Investment Behavior," *Amer. Econ. Rev.*, Sept. 1968, *58*, 681–712.

——— and ———, (1968b) "Optimal Capital Accumulation and Corporate Investment Behavior," *J. Polit. Econ.*, Nov./Dec. 1968, *76*, 1123–51.

H. Joseph, "Costs of Adjustment and the Flexible Accelerator," *Western Econ. J.*, Mar. 1970, *8*, 59–337.

M. Kendall and A. Stuart, *The Advanced Theory of Statistics*, Vol. 2, London 1961.

E. Kuh, *Capital Stock Growth: A Micro-Econometric Approach*, Amsterdam 1963.

T. Lancaster, "Business Saving and Normal Income," *Rev. Econ. Stud.*, Oct. 1963, *30*, 203–16.

D. Lindley, "Regression Lines and the Linear Functional Relationship," *J. Royal Statist. Soc.*, 1947, Suppl., 218–44.

R. Lucas, "Optimal Investment Policy and the Flexible Accelerator," *Int. Econ. Rev.*, Feb. 1967, *8*, 78–85.

J. Meyer and E. Kuh, *The Investment Decision*, Cambridge 1959.

M. Miller and F. Modigliani, "Some Estimates of the Cost of Capital to the Electric Utility Industry, 1954–57," *Amer. Econ. Rev.*, June 1966, *56*, 333–91.

E. Mills, *Price, Output and Inventory Policy*, New York 1962.

M. Nerlove, *The Dynamics of Supply: Estimation of Farmer's Response to Price*, Baltimore 1958.

G. Stevens, "On the Impact of Uncertainty on the Value and Investment of the Neoclassical Firm," *Amer. Econ. Rev.*, June 1974, *64*, 319–36.

J. Tinbergen, *Business Cycles in the United States, 1919–32*, Geneva 1939.

A. Treadway, "On Rational Entrepreneurial Behavior and the Demand for Investment," *Rev. Econ. Stud.*, Apr. 1969, *36*, 227–39.

# Rational Expectations in a Disequilibrium Model of the Term Structure

By MICHAEL E. ECHOLS AND JAN WALTER ELLIOTT*

Many of the recent tests of theories on the term structure of interest rates have been based on variants of an expectations model first developed by David Meiselman. Studies by Reuben Kessel and James Van Horne use variants of Meiselman's model in the attempt to empirically test various aspects of term structure theory. The models of Meiselman, Kessel, and Van Horne share the presumption that interest rate expectations are autoregressively related to prior and present components of the term structure according to an error-learning hypothesis.[1] Other work, most notably by Richard Roll and Thomas Sargent (1973), has examined the related proposition that interest rate expectations are a martingale or a sub-martingale. Sargent's results do not support the representation of interest rate expectations as a martingale or sub-martingale, while Roll's somewhat positive results are confined to the very short expectations embodied in the Treasury Bill market.

The formulation of the term structure as an essentially autoregressive model places the resulting theory in isolation from other components of the economic system and leaves several important questions unresolved. The role played by current economic variables in the formation of the type of interest rate expectations postulated by Irving Fisher, John Hicks, Friedrich Lutz, and others, remains unclear. Neither classical nor Keynesian concepts of interest rate determination are directly incorporated in the investor's expectational framework. Nonetheless, empirical evidence such as that presented by Sargent (1968), Martin Feldstein and Otto Eckstein, Thomas Yohe and Dennis Karnosky has shown that economic variables significantly affect the level of interest rates. Finally, the decreased ability of the error-learning model to explain the level of forward rates as longer time horizons are considered indicates the desirability of an expanded theory of investor expectations.[2]

In this paper, we develop a model of the term structure based on the premise that expectations of future spot rates are primarily determined by current and past economic behavior. The model is based upon a four-sector loanable funds representation of equilibrium interest rates. The solved structure of this model is used as the basis for the estimation of expected rates according to the rational expecta-

* Assistant professor and associate professor, respectively, School of Business Administration, University of Wisconsin-Milwaukee. We wish to acknowledge the financial support of the Northwestern Mutual Life Insurance Company, the Allis Chalmers Foundation, the UWM Graduate School, and the Management Research Center of the School of Business Administration.

[1] Meiselman's original structure is:

$$_{t+n}E_t - {}_{t+n}E_{t-1} = \beta(R_{t\cdot 1} - {}_tE_{t-1}) + \alpha$$

where $_{t+n}E_t$ is the expectation of the one-period spot rate $n$ periods in the future held by investors at time t; $R_{t\cdot 1}$ is the one-period spot rate occurring at time t; and $\alpha$, $\beta$ are constants.

[2] Meiselman's results for the error-learning model are:

| Year of Forward Rate | Correlation Coefficient |
|---|---|
| $n=1$ | .952 |
| 2 | .867 |
| 3 | .768 |
| 4 | .682 |
| 5 | .642 |
| 6 | .625 |
| 7 | .631 |
| 8 | .590 |

The error-learning model does an increasingly poor job of explaining longer forward rates.

tions concept proposed by John Muth. Our model of rationally based expected rates is incorporated into a disequilibrium framework that enables empirical testing of the role of expectations, liquidity factors, and institutional supply and demand factors in explaining fluctuations in forward rates over time.

In the pure expectations theory of the term structure discussed by Lutz and Meiselman, speculation is active across all maturity regions. The existence of risk-neutral investors drives the structure of forward rates implicit in the yield curve into close alignment with expected future spot rates. If forward rates rise above market expectations of future spot rates, demand builds for the appropriate forward contracts, bond prices rise for the maturities involved, and market equilibrium in which forward rates align with expected future spot rates is restored. Hicks and others have proposed modifications of this structure to accommodate a risk premium for longer maturity securities. In this view, observed forward rates consist of a risk premium as well as expectations of future spot rates. John Culbertson has argued that the existence of institutional barriers reduces the influence of speculative activity by inhibiting the leveling of forward rates with expectations. He proposes that customary investment maturity regions exist for a number of institutional investors, such that their liquidity requirements and cash flow position can affect some regions of the yield curve disproportionately more than other regions.

Empirical tests of the existence of risk premiums have been made by Meiselman, Van Horne, J. Huston McCulloch, and ourselves (1975a) using variants on the error-learning model described in footnote 1. Although Meiselman's original results are not sufficient to support his conclusions, for reasons given by Reuben Kessel, the conclusion of the other studies is that the existence of risk premiums appears to have positive empirical support. Direct empirical tests of the market segmentation ideas of Culbertson have been limited. Franco Modigliani and Richard Sutch (1966, 1967) have developed and tested a variant of Culbertson's model that emphasizes the impact of bond market supply upon market equilibrium between long and short maturity bonds. However, the results of their tests are quite mixed. In addition, Douglas Fisher has reported evidence of market segmentation for British securities, while Jacob Michaelson has conducted tests on holding period yields using Treasury data and has reported results unfavorable to a market segmentation view. Previous work undertaken by the present authors (see Elliott and Echols 1976) finds evidence of frequent market segmentation in longer maturity regions for U.S. Treasury bonds.

## I. A Disequilibrium Structure

None of the empirical tests now discussed has proceeded on the assumption that present expectations of future spot rates are primarily governed by aggregate economic behavior. By contrast, this assumption is a central premise of the work reported upon here. We begin with the well-accepted premise that forward market rates of interest contain substantial if not exclusive information on expectations of investors concerning future spot rates of interest. Forward rates are implied by the current term structure of yields according to:

$$(1) \qquad {}_{t+n}r_t \equiv \frac{(1 + R_{n+1,t})^{n+1}}{(1 + R_{n,t})^n} - 1$$

where $R_{n,t}$ is the spot rate of discount on $n$-period money as of time t and ${}_{t+n}r_t$ is the forward market rate of interest at time t on one-period money t$+n$ periods into the future that investors could be assured of by shorting $n$-period contracts and buying an

equal long position in $n+1$ period contracts.

To structure our model of forward rate determination, we employ a market disequilibrium framework.[3] This framework recognizes that at any point in time a given forward rate may not be fully adjusted to its equilibrium value, and that a given forward rate may or may not contain a liquidity premium. The disequilibrium framework is:

$$(2) \quad {}_{t+n}r_t \equiv {}_{t+n}E_t + ({}_{t+n}far_t - {}_{t+n}E_t) + ({}_{t+n}r_t - {}_{t+n}far_t)$$

where ${}_{t+n}E_t$ is the liquidity neutral fully adjusted expected spot rate for one-period money $n$ periods in the future held by investors at time t, and ${}_{t+n}far_t$ is the fully adjusted forward rate existing at time t for one-period money $n$ periods in the future.

The term $({}_{t+n}far_t - {}_{t+n}E_t)$ in equation (2) defines a liquidity premium term which is the difference between a fully adjusted forward rate and a fully adjusted liquidity neutral expected rate. The term $({}_{t+n}r_t - {}_{t+n}far_t)$ is the difference between the observed forward rate and the fully adjusted forward rate. This term defines a market equilibrium adjustment gap due to nonspeculative factors such as those proposed by Culbertson and Modigliani and Sutch.

In the pure expectations hypothesis, investors are neutral to liquidity considerations and markets are fully adjusted to equilibrium values. Thus, under that theory ${}_{t+n}r_t = {}_{t+n}E_t$ and both $({}_{t+n}far_t - {}_{t+n}E_t)$ and $({}_{t+n}r_t - {}_{t+n}far_t)$ are not significantly different from zero. In a fully adjusted liquidity premiums model, ${}_{t+n}r_t = {}_{t+n}E_t + ({}_{t+n}far_t - {}_{t+n}E_t)$ with $({}_{t+n}r_t - {}_{t+n}far_t)$ being essentially zero. Finally, if markets also lack effective short-term adjustment, the explanation of fluctuations in observed forward rates must involve all three terms.

### A. *Expectations*

As the first step in our analysis, we formulate a model structure to explain the expectations term $({}_{t+n}E_t)$. It is assumed that expectations of future interest rates are formed according to a rational expectations process as proposed by Muth where the economic determinants of current interest rate levels form the basis for investor's expectations about future actual rates.[4]

### B. *The Real Interest Rate*

The economic determinants of interest rate levels are identified by construction of a loanable funds model largely similiar to that proposed by Sargent (1968). In his model, nominal interest rates are composed of a real component, a component reflecting inflationary expectations, and a component measuring the efficiency of market equilibrium adjustments. For a given level of real output, the real interest rate adjusts so that aggregate flows of saving are offset by flows of new investment spending. Employing this concept, we follow Sargent in formulating equilibrium saving and investment functions as follows:

$$(3) \quad I(\mathrm{t}) = I(R_e(\mathrm{t}), L_1[\Delta X(\mathrm{t})]) \qquad \frac{\partial I}{\partial R_e} < 0; \quad \frac{\partial I}{\partial \Delta X} > 0$$

$$(4) \quad S(\mathrm{t}) = S(R_e(\mathrm{t}), X(\mathrm{t})) \qquad \frac{\partial S}{\partial R_e} > 0; \quad \frac{\partial S}{\partial X} > 0$$

[3] Our framework was inspired in part by a disequilibrium model proposed by Sargent (1968).

[4] A rational expectations process based upon economic determinants of the current interest rate is not inconsistent with either the expectations hypothesis or even Meiselman's model. In the error-learning model, the current one-period spot rate is implicitly determined by economic forces of supply and demand. Thus our model makes explicit the components of the implicit assumption contained in Meiselman's model that expectations are ultimately founded upon economic variables.

where:

$R_e(t)$ = the real interest rate
$I(t)$ = desired real investment
$S(t)$ = desired real saving
$X(t)$ = aggregate output
$\Delta X(t)$ = the one-period change in aggregate output
$L_1$ = lag operator

Expression (3) reflects a distributed lag accelerator similar to that studied by Robert Eisner, modified to account for interest rate effects on investment. Expression (4) is a simple Keynesian-type saving function. To these functions, we add an exogenous governmental sector including total spending $G(t)$ and taxes $T(t)$. We define the net budgetary position $B(t)$ as:

$$B(t) = G(t) - T(t) \tag{5}$$

We also add an exogenous net exports variable $NE(t)$. The condition for interest rate equilibrium is the equality of *ex ante* saving and investment, that is:

$$I(t) + B(t) + NE(t) - S(t) = 0 \tag{6}$$

Substituting expressions (3) and (4) into this framework allows solution for the equilibrium real interest rate $R_e^*(t)$ as:

$$R_e^*(t) = f(L_1[\Delta X(t)], X(t), B(t), NE(t)) \tag{7}$$

where the partial derivatives

$$\frac{\partial f}{\partial \Delta X} > 0; \quad \frac{\partial f}{\partial X} < 0; \quad \frac{\partial f}{\partial B} > 0; \quad \frac{\partial f}{\partial NE} > 0$$

evolve from the postulated signs of the saving and investment functions, and from the role of $B(t)$ and $NE(t)$ in the equilibrium expression (6).

### C. *Inflationary Expectations*

The impact of inflationary expectations on nominal interest rates is taken to be a distributed lag function of the rate of actual price change as postulated by Irving Fisher. This formulation has also been examined empirically by William Gibson, Leonall Anderson and Keith Carlson, Yohe and Karnosky, Elliott, and John Rutledge. Sargent (1973b) has noted that a particular set of restrictions on an *IS-LM* type model is implied when the Fisherian formulation of inflationary expectations is coupled with the assumption that $R_e^*$ is constant over time. Our formulation does not make this assumption and should be free of problems on this account.

The speed of adjustment component of nominal rates is assumed to depend principally upon changes in the money supply as proposed by Wicksell. When these two components are added to the real interest rate expression (7), the complete model for current nominal interest rates $i(t)$ results. This is:

$$i(t) = f(L_1[\Delta X(t)], X(t), B(t), NE(t)) + g(L_2[Dp(t)]) + h(DMS(t)) \tag{8}$$

where $Dp(t)$ is the percent change in general prices, $DMS(t)$ is the percent change in money supply, and $L_2$ is a second lag operator. Expression (8) differs from the loanable funds model proposed by Sargent in two ways. First, a distributed lag investment accelerator replaces his single period accelerator. This modification reflects the prevalent empirical evidence that output adjustments impact on investment demand with a lag. Secondly, government expenditures and net exports have been added to complete the model and to make it correspond to the recent economic conditions reflected in the modern data used here. Sargent ignored government spending, tax, and net export impacts on equilibrium in his tests using data from 1902–40, presumably because such effects were judged to be of little significance during this earlier period.

The complete model for expected spot

rates is produced from equation (8) by specifying the structure to be linear, defining the lag distributions, and by replacing $i(t)$ by ${}_{t+n}E_t$ to reflect the hypothesis that the economic factors impacting upon current nominal rates are the same factors that form rational expectations about future nominal rates.

In the empirical tests to follow, the use of monthly data has an impact on the specification of the two lag distributions. Since high serial correlation is commonly a characteristic of monthly data, even for flexibly constrained lag distributions of the considerable length suggested by previous studies of accelerator and price expectations lags, we formulated both of the lag distributions in the following way:

$$(9)\quad L_1 = a_0[X(t) - X(t-1)] + a_1\Delta X(1,12) + a_2\Delta X(13,24) + a_3\Delta X(25,36) + a_4\Delta X(37,48) + a_5\Delta X(49,60)$$

$$(10)\quad L_2 = \frac{b_0[p(t) - p(t-1)]}{p(t-1)} + b_1 Dp(1,12) + b_2 Dp(13,24) + b_3 Dp(25,36) + b_4 Dp(37,48) + b_5 Dp(49,60)$$

where the coefficients $a_i$ and $b_i$ $(i=1, 2, \ldots, 5)$ are unconstrained, and where

$$Dp(j, j+11) = \frac{p(t-j) - p(t-j-11)}{p(t-j-11)} \quad \text{for } j = 1, 13, 25, 37, 49$$

$$\Delta X(j, j+11) = X(t-j) - X(t-j-11) \quad \text{for } j = 1, 13, 25, 37, 49$$

We may now write the model of expected future spot rates used in our empirical testing:

$$(11)\quad {}_{t+n}E_t = C_0 + C_1X(t) + C_2B(t) + C_3NE(t) + C_4DMS(t) + L_1[\Delta X(t)] + L_2[Dp(t)]$$

D. *Liquidity and Market Disequilibrium*

Our model of observed forward rates is completed by specifying determinants of the liquidity and market disequilibrium terms. Earlier work by Kessel, Van Horne, and Charles Nelson has utilized the idea that liquidity premiums are influenced by temporal changes in market interest rates or other related variables.[5] In particular, Kessel presents supporting empirical evidence that liquidity premiums are a function of current spot rates on short maturity default free debt. We test this hypothesis using our model of expectations by specifying liquidity premiums to be a function of the current one-year spot rate on *U.S.* government bonds. Thus, the first specification of the liquidity premium term is that it is proportional to current spot rates:

$$(12)\quad ({}_{t+n}far_t - {}_{t+n}E_t) = C_5(R_{t,1}); \quad C_5 \geq 0$$

where $R_{t,1}$ is the current one-year spot rate.

Van Horne has conducted empirical tests related to liquidity premiums using the error-learning framework and focusing upon forward rates rather than spot rates. He interprets his results as indicating the importance of interest rate risk, or the risk associated with fluctuations in the principal value of securities due to changes in interest rate levels.[6] Van Horne finds this risk to be inversely related to forward interest rate levels. Increases in bond prices relative to an accustomed price are assumed to increase interest rate risk essentially by increasing the probability of future price decreases. To test this hypothesis in

[5] A direct measurement of a liquidity premiums function in the context of the error-learning model has been presented in a separate work by the authors (1975).

[6] The conceptual development of interest rate risk is presented by Nicholas Kaldor and Burton Malkiel. Charles Nelson, pp. 35–36, cautions that the empirical results presented by Van Horne in his test of an interest rate risk variable are suspect due to the existence of a specification error in the original model. Further research on Van Horne's original model and extended empirical results on the interest rate risk variable are available in the authors' paper (1975).

our analysis, the liquidity premium is taken to be a function of the price of bonds having one year to maturity. That is:

$$(13)\quad ({}_{t+n}far_t - {}_{t+n}E_t) = C_6(P_{t,1});\quad C_6 \geq 0$$

The market adjustment term in expression (2) focuses upon the extent to which speculative activity maintains markets in a near-equilibrium position over time. Culbertson's work on this subject has focused on motives for investment by banks, insurance companies, and other financial institutions that potentially restrict their speculative activity to certain regions of the yield curve. More recently, Modigliani and Sutch (1967, 1968) have been concerned with the possible impact on the yield curve of changes in the proportion of *U.S.* government debt of different maturities supplied to the investment markets. In Culbertson's view, the demand for bonds by particular institutions can be expected to affect some regions of the yield curve disproportionately without such effects being transmitted along the entire curve.[7] Modigliani and Sutch point to the same type of rigidities in acceptable maturity regions, i.e., preferred habitats, as potential reasons to expect the supply of bonds of a particular maturity to push up yields in the associated maturity region by a disproportionate amount.

To obtain an approximate measure of the movements of institutional patterns of demand, we calculate a ratio of the stock of investment funds available to banks compared to that available to insurance companies. These two institutions are typically both large holders of *U.S.* government securities. Furthermore, data regularly published in the *Federal Reserve Bulletin* (see Table A37–Dec. 1974 issue for example) reveal that bank holdings of *U.S.* government securities are concentrated in predominantly short-maturity issues while insurance companies hold disproportionately large amounts of long-maturity issues. In markets that adjust quickly and completely, fluctuations in the ratio of bank to insurance company holdings should have no impact on forward market rates as the demand and supply effects will be transmitted along the entire yield curve. The specific hypothesis tested here is:

$$(14)\quad ({}_{t+n}r_t - {}_{t+n}far_t) = C_7\left[\frac{BH(\mathrm{t})}{IH(\mathrm{t})}\right]$$

where $BH(\mathrm{t})$ is the stock of bank funds invested in *U.S.* government securities at time t and $IH(\mathrm{t})$ is the stock of insurance company funds invested in *U.S.* government securities at t.

In measuring the movement in the relative supplies of debt of various maturities, we use the approach of Modigliani and Sutch by computing the proportion of debt in various maturity categories. They found the ratio of short to long debt to be among the best of a number of candidate measures of bond supply effects. We define such a measure as the ratio of the stock of bonds having less than 5 years to maturity to the stock of bonds having 10 years or greater to maturity. The split at 5 years is inspired by the availability of data and previous work by ourselves (Elliott and Echols 1976) in which the 8-year maturity region for *U.S.* government securities was found to exhibit a clear break in the speculative characteristics between the short and long regions. In a highly speculative market for governments, changes in the proportion of short to long bonds outstanding should have no explanatory power relative to forward rate movements. We test this proposition using the following hypothesis:

$$(15)\quad ({}_{t+n}r_t - {}_{t+n}far_t) = C_8\cdot\left(\frac{SG(\mathrm{t})}{LG(\mathrm{t})}\right)$$

[7] While Culbertson discusses the demand effects, Modigliani and Sutch (1966) assume these effects are unimportant as follows: "Unfortunately, the measurement of supply-demand effects poses formidable problems, even if we are prepared to limit ourselves to variations in supply, on the assumption that the demand side is not subject to significant variations" (p. 191).

where $SG(t)$ is the stock of 0–5 years to maturity $U.S.$ government bonds at t and $LG(t)$ is the stock of 10-year and up $U.S.$ government bonds at time t.

To summarize, we have now formulated an explanatory model for forward rates from expression (2) that contains as arguments rationally expected future rates (expression (11)) and may also contain liquidity influences (expressions (12)–(13)) and/or supply and demand influences (expressions (14)–(15)).

## II. Data and Empirical Results

The empirical tests of the model presented in this paper are based on month-ending yield data derived from a yield curve model developed by the authors (1975) and presented in the Appendix. The yields are produced by setting the coupon variable in the yield curve model equal to zero and computing the zero-coupon-equivalent yield for each maturity tested.[8] Details of this and other calculations are given in the Appendix.

The yield curve values enable the dependent variable in our analysis, forward rates, to be derived according to equation (1). The existence of a seasonal pattern motivated the seasonal adjustment of the raw forward rates.

An important attribute of the yield and derived forward-rate data used here is its correction for coupon effects. The correct calculation of forward rates must be based on pure rates of discount, rather than from coupon-bearing bonds, basically due to the return of capital through time associated with coupon-bearing bonds. Thus, ideally, forward rates should be calculated from yields on zero-coupon bonds. In our data, we approximate this pure discount rate by our treatment of the coupon term. The forward rates that result are free of distortions that may result from the confounding of coupon effects upon yield with pure term structure effects.[9]

The initial empirical test is conducted upon the rational interest rate expectations model component of expression (11), where the observable forward rate ${}_{t+n}r_t$ replaces the unobservable ${}_{t+n}E_t$. In the statistical estimation of (11), it is important to note that the observation on the dependent variable ${}_{t+n}r_t$ is taken at the close of trading for month t while observations on all t subscripted right-hand-side variables are taken as flows over period t or as average stocks during t. Thus, no opportunity for "feedback" of a simultaneous-equations nature is present in our results, and all same-period right-hand-side variables can be taken as predetermined with respect to the dependent variable. Behaviorally, taking the observations on ${}_{t+n}r_t$ at month end postulates that the month-ending term structure adjusts so that the cumulative flows of aggregate saving and investment during the month are brought into balance. This approach seems as reasonable as postulating an equivalent statement about month-averaged forward rates, and is clearly superior with respect to the econometric conveniences introduced by it.

The expectations model is fit for each of the ten forward rates, with the results shown in Table 1. These results show a generally high degree of forward rate explanation implying our rational expectations model has useful explanatory properties. In addition, the corrected $R^2$ values increase with more distant forward rates, revealing that the expectations model explains a somewhat larger proportion of the fluctuation in forward rates as those rates become more distant.[10] This pattern is in-

[8] The range of observed coupon values in our yield data is from $2\frac{1}{2}$ percent to 7 percent.

[9] Previous empirical studies that have used David Durand's data or U.S. Treasury yield data are subject to the difficulty associated with intermixing bonds of various coupons and rates in samples used to calculate yield curves.

[10] This is not merely the result of differences in the

TABLE 1—RATIONAL EXPECTATIONS MODEL

| Maturity (years) | $C_0$ | $C_1$[a] | $\sum a_{it}$[b] | $C_4$[c] | $\sum b_i$[d] | $C_3$[e] | $C_2$[f] | Corrected $R^2$ | $D.W.$ | Standard Error |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.08918 | −0.0944 | 1.3312 | 0.0026 | 0.8807 | −1.2233 | 2.9332 | .872 | 1.22 | .0040 |
| | (2.10) | (0.91) | (0.40) | (0.02) | (1.89) | (−0.31) | (2.82) | | | |
| 2 | 0.09124 | −1.0246 | 1.8367 | −0.0059 | 0.9586 | −0.8910 | 2.6665 | .878 | 1.22 | .0038 |
| | (2.27) | (1.04) | (0.58) | (0.05) | (2.18) | (−0.24) | (2.71) | | | |
| 3 | 0.0924 | −1.0912 | 2.3650 | −0.0182 | 1.0331 | −0.4131 | 2.3956 | .885 | 1.21 | .0036 |
| | (2.42) | (1.17) | (0.79) | (0.16) | (2.48) | (−0.12) | (2.57) | | | |
| 4 | 0.0927 | −1.1341 | 2.7467 | −0.0308 | 1.0919 | 0.0620 | 2.1290 | .891 | 1.21 | .0034 |
| | (2.57) | (1.29) | (0.96) | (0.28) | (2.76) | (0.02) | (2.41) | | | |
| 5 | 0.0934 | −1.1777 | 3.0333 | −0.0373 | 1.1434 | 0.3875 | 1.8846 | .899 | 1.21 | .0032 |
| | (2.75) | (1.12) | (1.13) | (0.36) | (3.07) | (0.12) | (2.27) | | | |
| 7 | 0.0900 | −1.1520 | 3.0460 | −0.0518 | 1.1823 | 1.3980 | 1.4370 | .913 | 1.22 | .0029 |
| | (2.98) | (1.56) | (1.27) | (0.56) | (3.58) | (0.51) | (1.95) | | | |
| 9 | 0.0829 | −1.0442 | 2.6566 | −0.0730 | 1.1766 | 2.8247 | 1.0389 | .927 | 1.26 | .0026 |
| | (3.08) | (1.59) | (1.25) | (0.89) | (3.99) | (1.15) | (1.58) | | | |
| 12 | 0.0833 | −1.1431 | 3.3979 | −0.0796 | 1.3195 | 4.1646 | .3869 | .944 | 1.41 | .0022 |
| | (3.62) | (2.04) | (2.00) | (1.13) | (5.24) | (1.98) | (0.69) | | | |
| 15 | 0.0899 | −1.3972 | 5.0125 | −0.0772 | 1.3756 | 5.1805 | 0.2985 | 952 | 1.66 | .0021 |
| | (4.16) | (2.65) | (3.26) | (1.17) | (5.82) | (2.63) | (0.57) | | | |
| 20 | 0.1018 | −1.8063 | 7.1418 | −0.0686 | 1.9244 | 6.4982 | −1.3571 | .945 | 1.72 | .0024 |
| | (4.07) | (2.96) | (3.35) | (0.90) | (7.03) | (2.84) | (2.22) | | | |

*Note:* $R^2$ values are the "corrected" statistic throughout. Numbers in parentheses are $t$-statistics, while for lag distributions they are the $t$-statistic associated with the summed value of the distribution and calculated from the pooled variance.

[a] Real output $X(t)$

[b] Change in real output $L_1[\Delta X(t)]$

[c] Change in money supply $DMS(t)$

[d] Price change $L_2[Dp(t)]$

[e] Net exports $NE(t)$

[f] Government deficit $B(t)$

terpreted to suggest that expectations about future spot rates play a slightly more significant role in forward rate determination the more distant the forward rate. At first glance this result seems the opposite of that developed by Meiselman, who found the fit of the error-learning model to decrease over the maturity range from 1 to 8 years. However, the two are not necessarily incompatible. Meiselman has shown that error learning is less important for more distant rates; more important for closer rates. Our results suggest rational expectations are slightly more important for more distant rates; less important for closer rates. The two sets of results are not inconsistent with the suggestion that more weight passes from error-learning influences to rational expectations influences as more distant rate expectations are formulated.

The variables in our expectations model generally produced measured coefficients that are consistent with their a priori role in the model. We find real output to have a uniformly negative effect on interest rate expectations reflecting its role in the saving function. The governmental deficit is found to have the expected positive effect in all but the two longest maturities, and the investment accelerator $L_1[\Delta X(t)]$ and price expectations $L_2[Dp(t)]$ terms are found to each have positive effects as hypothesized. The impact of net exports on interest rate expectations are found to be weakly negative for the first three forward rates and increasingly positive thereafter. This may reflect the initial monetary effect of a change in net exports, which is expected to be negative, followed by the longer-run expectation of rising rates due to the increase in investment-like expenditures and the attendant positive impact on forward rates.

The money supply term is negative in

total variability or mean value of the various forward rate series, for equation (A1) of the Appendix shows both the mean and the standard deviation to be approximately constant over the entire forward rate sample.

TABLE 2—LIQUIDITY TERMS

| | One-Year Spot Rates | | | | | One-Year Bond Prices | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Year | Value of $C_5$ | $R^2$ | $D.W.$ | $F$-value | Standard Error | Value of $C_6$ | $R^2$ | $D.W.$ | $F$-value | Standard Error |
| 1 | .0064 | .922 | 1.78 | 58.8[a] | .0032 | −0.7104 | .921 | 1.78 | 58.0[a] | .0032 |
| | (7.65) | | | | | (7.60) | | | | |
| 2 | .0064 | .925 | 1.76 | 57.3[a] | .0030 | −0.6667 | .930 | 1.75 | 56.7[a] | .0030 |
| | (7.55) | | | | | (7.49) | | | | |
| 3 | .0057 | .929 | 1.74 | 56.8[a] | .0029 | −0.6309 | .928 | 1.73 | 55.9[a] | .0029 |
| | (7.54) | | | | | (7.48) | | | | |
| 4 | .0054 | .933 | 1.70 | 55.5[a] | .0027 | −0.5943 | .932 | 1.70 | 54.6[a] | .0027 |
| | (7.49) | | | | | (7.43) | | | | |
| 5 | .0050 | .936 | 1.68 | 53.7[a] | .0026 | −0.5513 | .936 | 1.67 | 52.7[a] | .0026 |
| | (7.31) | | | | | (7.25) | | | | |
| 7 | .0043 | .944 | 1.62 | 49.3[a] | .0023 | −0.4781 | .944 | 1.61 | 48.3[a] | .0023 |
| | (7.05) | | | | | (6.98) | | | | |
| 9 | .0037 | .951 | 1.58 | 44.0[a] | .0021 | −0.4071 | .950 | 1.57 | 42.8[a] | .0021 |
| | (6.60) | | | | | (6.53) | | | | |
| 12 | .0025 | .955 | 1.57 | 23.3[a] | .0020 | −0.2760 | .955 | 1.57 | 23.3[a] | .0020 |
| | (4.83) | | | | | (4.76) | | | | |
| 15 | .0012 | .954 | 1.67 | 5.65 | .0020 | −0.1321 | .954 | 1.67 | 4.62 | .0020 |
| | (2.28) | | | | | (2.23) | | | | |
| 20 | −.0012 | .947 | 1.83 | 3.99 | .0023 | 0.1392 | .947 | 1.83 | 4.41 | .0023 |
| | (1.98) | | | | | (2.02) | | | | |

*Note*: Numbers in parentheses are $t$-values. See John Johnston, pp. 145–47 for the explanation of the $F$-test for the significance of the added variables.

[a] Significant at 1 percent level.

all but the initial period where it is positive and insignificantly large. This prevalent negative influence of current money supply change on expected interest rates is consistent with results obtained by Sargent (1968) in his empirical tests, and consistent generally with the theoretical role of money in the present model structure. In addition, while lacking in statistical significance, we find the role of the monetary effect to grow somewhat with more distant rate expectations. Taken together, these results suggest that the variables in our rational expectations model work individual effects that are consistent with their a priori roles in the loanable funds framework.

The second sets of tests explore the impact of adding the liquidity premium and institutional supply and demand terms to the expectations model. Table 2 shows the estimated coefficients and goodness-of-fit statistics associated with the separate addition to the interest rate expectations model structure of each of the two variables used to test for liquidity premiums. The re-estimates of the coefficients in the expectations model to which these terms are added are not shown to simplify the table. However, they do not differ measurably from the results shown in Table 1 with the exception discussed below.[11] Standard errors in Table 2 are directly comparable with those in Table 1 since the dependent variable is identical.

Both current 1-year spot rates and 1-year bond prices are found to have the sign proposed by Kessel with the exception of the 20-year rate, where the effect fades in significance. Thus, they are inconsistent with Van Horne's hypothesis. Increases in

[11] These details and similar details associated with Tables 2 and 3 can be provided to the interested reader upon request.

current spot rates or declines in bond prices are found here to systematically increase forward rates, presumably due to their role in influencing liquidity premiums. $F$-tests of the additional explanatory power associated with each term are shown in Table 2. These results show that either variable tested adds substantial explanatory power to that of the basic expectations model. This result is further support for Kessel's hypothesis on the role of current spot rates in determining liquidity premiums. However, it is not possible to assign the influence of current spot rates uniquely to liquidity premium effects, due to the possible reflection of expectational determinants in current short-term rates. In this respect, we find that the impact of the government deficit position changes signs and declines in significance when the spot rate or bond price is added to the equation. This indicates a collinear pattern between the two, and more broadly suggests that the same type of future interest rate expectational determinants represented by the deficit position are contained in current spot rates. Even so, the information contained in spot rates is sufficiently independent to cause this term to make a statistically significant contribution to the basic expectations model.

The magnitude of the impact of both the spot rate and bond price terms declines somewhat as the forward rate horizon is lengthened. This produces a suggestion that the liquidity related effects we have measured depend somewhat upon the extent to which rational expectations explain forward rate levels. At 1-year forward rates, when the rational expectations model explains 87.2 percent of the forward rate fluctuations, both the liquidity premium variables exert their largest effects, while at 20-year forward rates where rational expectations account for 94.5 percent of the fluctuation in forward rates, the liquidity terms exert their smallest and least significant effects. Thus, our results suggest the perception of a liquidity premium as a price paid for the unpredictability of future rates, i.e., as the markup over expectations that investors require due to the incompleteness with which rational expectations successfully account for movements in forward rates.

Finally, the spot rate term appears to be slightly better than the bond price in its explanatory power. The $F$-values for the addition of this variable are a little higher than for the bond price, while both have a nearly identical effect in reducing the somewhat questionable degree of autocorrelation associated with the basic expectations equation. In considering the market equilibrium variables, we proceed by using the spot rate to reflect liquidity effects on forward rates.

We consider the market equilibrium adjustment effect in two ways: in relation to the basic equation, and in relation to the basic equation with the spot rate term added. Although the results of these two tests lead to similar conclusions and interpretations, differing primarily in the magnitude of the $F$-statistics, there is reason to prefer the results from the the latter structure including the spot rate term. Our analysis to this point has suggested that both rational expectations and the liquidity premium variable play influential roles in forward rate determination. Given the nature of these variables, we can only interpret their role as influencing equilibrium positions. In considering possible influences of market adjustment imperfections, the most complete equilibrium model is the most appropriate. Thus, we consider the market adjustment variables in combination with the expectational and liquidity premiums portions of the model.

The results of adding the institutional demand and bond supply variables to the expectations and liquidity premium structure are shown in Table 3. The ratio of

TABLE 3—INSTITUTIONAL DEMAND AND BOND SUPPLY EFFECTS

| | Years | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 7 | 9 | 12 | 15 | 20 |
| **Institutional Demand**[a] | | | | | | | | | | |
| Value of $C_7$ | −2.2727 | −2.1991 | −2.0441 | −1.9390 | −1.8286 | −1.5486 | −1.2909 | −0.9198 | −0.5888 | 0.1347 |
| | (3.09) | (3.15) | (3.08) | (3.08) | (3.05) | (2.85) | (2.59) | (1.96) | (1.21) | (0.23) |
| $R^2$ | .929 | .932 | .935 | .938 | .942 | .948 | .954 | .957 | .954 | .946 |
| *D.W.* | 1.92 | 1.89 | 1.86 | 1.82 | 1.78 | 1.69 | 1.60 | 1.55 | 1.64 | 1.84 |
| *F*-Value | 9.46[c] | 10.04[c] | 9.52[c] | 9.51[c] | 9.33[c] | 8.52[c] | 6.18[c] | 4.41[d] | 0.50 | 0.16 |
| Standard Error | .0030 | .0029 | .0027 | .0026 | .0025 | .0022 | .0020 | .0019 | .0020 | .0024 |
| **Bond Supply**[b] | | | | | | | | | | |
| Value of $C_8$ | 6.4040 | 6.2883 | 5.0029 | 4.6008 | 4.1156 | 2.3395 | 3.9114 | −2.1600 | −4.1909 | −7.8085 |
| | (0.66) | (0.68) | (0.57) | (0.54) | (0.52) | (0.33) | (0.06) | (0.36) | (0.68) | (1.10) |
| $R^2$ | .921 | .925 | .928 | .932 | .936 | .943 | .950 | .955 | .954 | .947 |
| *D.W.* | 1.79 | 1.77 | 1.74 | 1.71 | 1.68 | 1.62 | 1.58 | 1.58 | 1.69 | 1.87 |
| *F*-Value | 0.51 | 0.53 | 0.44 | 0.26 | 0.14 | 0.13 | 0.00 | 0.47 | 0.20 | 1.34 |
| Standard Error | .0032 | .0030 | .0029 | .0027 | .0026 | .0023 | .0021 | .0020 | .0020 | .0023 |
| **Demand and Supply** | | | | | | | | | | |
| Value of $C_7$ | −2.2699 | −2.1946 | −2.0555 | −1.9560 | −1.8454 | −1.5841 | −1.3494 | −1.0066 | −0.6963 | − .0030 |
| | (2.99) | (3.05) | (3.00) | (3.01) | (2.99) | (2.83) | (2.63) | (2.08) | (1.39) | (.01) |
| Value of $C_8$ | .1618 | .2529 | −.6499 | −.9185 | −.9575 | −2.0168 | −3.3197 | −4.9283 | −6.1058 | −7.8167 |
| | (0.02) | (0.03) | (0.08) | (0.11) | (0.12) | (0.29) | (0.52) | (0.82) | (0.98) | (1.06) |
| $R^2$ | .928 | .931 | .934 | .938 | .941 | .948 | .953 | .957 | .954 | .946 |
| *D.W.* | 1.92 | 1.89 | 1.86 | 1.82 | 1.78 | 1.69 | 1.60 | 1.56 | 1.67 | 1.87 |
| *F*-Value | 4.68[d] | 5.19[d] | 4.71[d] | 4.73[d] | 4.64[d] | 4.26[d] | 3.32[d] | 2.36 | 0.74 | 0.66 |
| Standard Error | .0030 | .0029 | .0027 | .0026 | .0025 | .0022 | .0021 | .0019 | .0020 | .0024 |

*Note*: Values in parentheses are *t*-values.

[a] $\left[\dfrac{BH(t)}{IH(t)}\right]$

[b] $\left[\dfrac{SG(t)}{LG(t)}\right]$

[c] Significant at the 1 percent level
[d] Significant at the 5 percent level

bank to insurance company holdings shows the expected negative sign in all cases but the 20-year case where the relationship is measured to be weak and in the opposite direction. The indication is that an increase in the proportion of investment funds held by banks relative to insurance companies systematically pushes down forward market rates, presumably due to their short preferred habitat. Furthermore, the statistically significant $F$-values associated with addition of the bank to insurance company ratio indicate that explanatory significance attaches to fluctuations in this ratio that are not contained in either the expectations model or the liquidity term. In addition, the value of $C_7$ declines in impact as forward rates become more distant, indicating that positive shifts in the proportion of bank investment funds has a somewhat greater impact on closer forward rates than on more distant forward rates.

The ratio of short to long maturity bond supplies also shows a measured sign consistent with its a priori role in all but the longest maturity cases. Thus an increase in the proportion of short maturity debt outstanding has a small positive impact on forward market rates, increasingly for more current forward rates than for more distant rates. However, this term in no case makes a statistically significant contribution to the explanation of forward rate movements, as the $F$-values are insignificantly small. In fact, the bond supply term actually detracts from the level of additional explanation associated with the bank to insurance ratio when the two terms are added to the model as shown in the third set of results in Table 3. These results are in basic agreement with the

studies of bond supply made by Modigliani and Sutch in not finding a substantial explanatory role for such ratios. We conclude that shifts in institutional participation in bond markets between banks and insurance companies are significant in forward rate determination suggestive of the operations of preferred habitats for banks and/or insurance companies. However, evidence has not been obtained that supplies of short vs. long bonds have a significant impact on forward rates.

### III. Rational Interest Rate Expectations and Preferred Habitats

Our rational expectations model for expected future equilibrium interest rates given by expression (11) can be integrated with the basic model structure of Modigliani and Sutch to develop new evidence on the usefulness of their approach in explaining yield spreads, and on the potential role of the earlier defined bond supply and bank-insurance company holdings terms in contributing to the explanation of yield spreads. Modigliani and Sutch's basic model is a statement of the equality of holding period returns across various maturities. It is given as follows:

$$(16) \quad R_{n,t} = R_{m,t} - [\textit{Expected Capital Gain}] + F_t$$

where $R_{n,t}$ is the market yield on $n$-period money, where $m$ is the minimum holding period, and where $F_t$ is the net effect of relative supply factors. In Modigliani and Sutch's work, the expected capital gain term in (16) is assumed to be proportional to the expected fall in the yield on a present $n$-period bond over the given holding period of $m$ length. This gives:

$$(17) \quad R_{n,t} = R_{m,t} - \beta[R_{n,t} - {}_{t+m}E_{n-m,t}] + F_t$$

where ${}_{t+m}E_{n-m,t}$ is the rate on $n-m$ period money expected to exist $m$ periods in the future, held as of time t.

Modigliani and Sutch replace the terms in the bracket of expression (17) by a distributed lag on short rates. The development of the present rational expectations model for expected rates and the positive support we have obtained for its structure enables analysis of an alternative approach. We may instead insert a direct estimate of ${}_{t+m}E_{n-m,t}$ into (17) obtained by regressing an actual rate upon the right-hand-side determinants of the rational expectations generating function of expression (11) and computing the value of ${}_{t+m}E_{n-m,t}$ from the estimated coefficients.

Use of a direct estimate of the expected rate term in expression (17) leaves the $R_{n,t}$ term on both sides of the equation. Solving for this term and assuming $F_t$ to be proportional to various market disequilibrium measures $X_t$ according to a proportionality constant $\alpha$ gives:[12]

$$(18) \quad R_{n,t} = k_1 R_{m,t} + k_2 \cdot {}_{t+m}E_{n-m,t} + k_3 \cdot X_t$$

$$\text{where } k_1 = \frac{1}{1+\beta}, \quad k_2 = \frac{\beta}{1+\beta}, \quad k_3 = \frac{\alpha}{1+\beta}$$

To estimate (18) we must select the maturity $n$ and the holding period $m$. Modigliani and Sutch did most of their work using a sample of bond yields containing a mixture of securities in the 10 to 15-year maturity range for the dependent variable and with a variety of differing coupons. For consistency with their maturity, we use a 12-year government yield obtained from our zero-coupon equivalent data. We consider a 90-day holding period ($m=1/4$) corresponding to Modigliani and Sutch's holding period. Accordingly, in estimating coefficients for the calculating of ${}_{t+m}E_{n-m,t}$, we use the one-quarter ahead yield on a

[12] Inspection of the expressions for $k_1$, $k_2$, $k_3$ shows an implied constraint on the value of $\beta$ in these regressions. We have not imposed this constraint in our empirical tests in order to observe the extent to which the data exhibit the constraint implied by the model. We find the model constraint on $\beta$ is not well characterized by the data.

bond of 11 years 9 months to maturity for the left-hand side of expression (11). Similarly, $R_{m,t}$ is approximated by a 90-day Treasury Bill rate as in the Modigliani-Sutch study. Finally, to compare our results directly with those of Modigliani-Sutch, we subtract $R_{m,t}$ from both sides of the equation, leaving the dependent variable defined as a spread between the 12-year and 90-day rate, and redefining the coefficient $k_1$ in expression (18) to be $(k_1-1)=k_1'$ as is done in the Modigliani-Sutch estimates. Thus, we estimate:

$$(19) \quad S_t = k_1' R_{m,t} + k_2 \cdot {}_{t+m}E_{n-m,t} + k_3 \cdot X_t$$

where $S_t$ is the spread.

To measure the effects of bond supply, we use the earlier defined ratio of short to long *U.S.* government bonds outstanding, and in addition we follow Modigliani and Sutch in the use of the proportion of short debt (5 years and under) to total debt and the proportion of long debt (10 years and over) to total debt. To these three representations of their bond supply measures, we add parallel measures of institutional participation due to our earlier results: (a) the ratio of bank investment funds to total debt, (b) the ratio of insurance company investment to total debt, and (c) the earlier-defined ratio of bank-to-insurance company investment funds.

These measures are combined into three alternative measures of the $F_t$ influence in expression (16):

$$\text{(Short Funds)} \quad F_t = k_{3,D} \cdot \frac{DH(t)}{T(t)} + k_{3,S} \cdot \frac{SG(t)}{T(t)}$$

$$\text{(Long Funds)} \quad F_t = k_{3,D} \cdot \frac{IH(t)}{T(t)} + k_{3,S} \cdot \frac{LG(t)}{T(t)}$$

$$\text{(Short-Long Ratios)} \quad F_t = k_{3,D} \cdot \frac{BH(t)}{IH(t)} + k_{3,S} \cdot \frac{SG(t)}{LG(t)}$$

where $T(t)$ is the total outstanding *U.S.* government securities and other variables are as defined earlier. The results of estimating these three equations plus the basic equation where $F_t=0$ are given in Table 4. In these estimations, an intercept term $k_0$ has been added to provide a weak test of this aspect of the model specification. In no case was it found to be significantly different from zero in magnitude, a result that is consistent with the model structure of expression (19). The estimated values of $k_2$ provide strong support for the role of the expected future rate variable with which it is associated, as this coefficient is highly significant in each of the model specifications in which it appears. In addition, the impact of the short-term

TABLE 4—MODIGLIANI-SUTCH MODEL

| | Coefficients of Expression (19) | | | | | | |
|---|---|---|---|---|---|---|---|
| | $k_0$ | $k_1'$ | $k_2$ | $k_{3,D}$ | $k_{3,S}$ | $R^2/D.W.$ | $S.E./F$ |
| Basic | .0015 | −.7690 | .7264 | — | — | .479/2.10 | 3.08 |
| | (0.39) | (9.90) | (7.35) | | | | |
| Short | −.0290 | −.7552 | .5455 | .0001 | .0501 | .520/2.27 | 2.95/5.49[a] |
| Funds | (1.20) | (10.06) | (4.42) | (0.00) | (2.98) | | |
| Long | .0128 | −.6342 | .5450 | .0008 | −.2428 | .525/2.33 | 2.94/6.08[a] |
| Funds | (1.30) | (7.58) | (4.36) | (2.11) | (3.47) | | |
| Short-Long | .0075 | −.6571 | .4639 | −1.2017 | .0021 | .541/2.32 | 2.89/8.14[a] |
| Ratios | (1.20) | (8.32) | (3.88) | (2.19) | (3.99) | | |

*Note:* Equations are estimated by generalized least squares with $\rho=.60$ due to the highly autoregressive nature of the ordinary least squares (*OLS*) error term. Numbers in parentheses are computed $t$-values. Standard errors are given in ($10^{-3}$) units of the dependent variable. $F$-values are for the significance of adding the $k_{3,D}$ and $k_{3,S}$ terms.

[a] Significant at the 1 percent level.

rate on the spread is negative as hypothesized and statistically significant in each equation. This is also consistent with the estimates of Modigliani-Sutch (1966). Furthermore, the magnitude of the estimated coefficients is quite comparable in the two cases ranging from about $-.65$ to $-.77$ in our case to about $-.69$ in the Modigliani-Sutch case.

The $F$-values given in Table 4 show the significance of the increase in explained variation associated with adding the alternative specifications of the $F_t$ influences to the basic equation. For all three specifications, the $F$-values are significant beyond the 1 percent level, indicating the supply and demand measures have substantial explanatory significance. The long-short ratio model provides the greatest effect in this regard, indicating this form of the variables to be somewhat superior. However, the measured direction of effect associated with each supply and demand variable is in the wrong direction vis-à-vis its underlying hypothesized role. This result is quite similiar to that found by Modigliani and Sutch (1966, p. 191); indeed our coefficient estimates are in the same direction as comparable estimates of theirs in all but the case of the supply of long bonds variable.

In interpreting the original Modigliani-Sutch results, it might be suggested that the omission of measures of institutional demand contributed to the peculiar results obtained for the supply variables; however, we find here that while the demand variables used are significant, their presence does not alter the result obtained for the supply ratios. Thus, it is necessary to look elsewhere to find the explanation for the conflict between the a priori role and the empirical results for the supply terms.

## IV. Conclusion and Implications

An explanatory model for forward rates has been developed in this research that enables analysis of the contribution of expectational determinants, liquidity premium influences, and institutional demand and bond market supply effects upon fluctuations in forward rates. A rational expectations model for future expected interest rates has been developed from a loanable funds framework. The results of empirical tests of this rational expectations structure show it to have consistently high $R^2$ values and individual coefficient signs that are consistent with the a priori rôle of each variable in the underlying loanable funds model of interest rate determination. These results are supportive of the relevance of the model in measuring rational interest rate expectations.

When spot rates are added to the expectations model structure to measure liquidity premium effects, the degree to which fluctuations in forward rates are explained is significantly increased. The additional expansion of the model to include the ratio of bank to insurance company investment funds further significantly increases the explanation of forward rate fluctuations. However, the explanatory role of relative bond supplies in forward rate explanation is not confirmed in a similar experiment. Taken together these results suggest an important incremental role for a liquidity-premiums–related measure and a measure of institutional bond demand in affecting forward rate fluctuations independent of the significant role associated with rationally expected future rates.

The present model of rationally expected future interest rates is used in the model structure of Modigliani and Sutch to obtain new evidence on the impact of institutional demand and bond supply upon yield spreads. Unlike the results obtained in our model of forward rates, empirical evidence with respect to fluctuations in the yield spread between 12-year government securities and the Treasury Bill rate is not consistent with the a priori

role assigned to these variables.

One possibility for this divergence in logical versus empirical results is that yield spreads are not a sufficiently sensitive vehicle for observing such effects. While yield spreads show the total change in yield over a given maturity range, forward rates approximate the movement in the slope of the yield curve over particular short regions of the curve. Thus, our forward rate model results could be taken to suggest that institutional demand patterns influence the instantaneous slope of the yield curve at a number of points along the curve. At the same time, the tests of the Modigliani-Sutch model indicate that raw yield spreads over a fixed maturity range are not similarly influenced.

We have not produced evidence that suggests that specific debt management strategies of supplying bonds of particular maturities are likely to influence either spreads or forward rates. However, our results support the suggestion that policy actions that alter the relative amounts of funds available to institutions such as banks and insurance companies might well be expected to impact on forward rates and thus upon the relative shape of the yield curve. Thus, policy actions having the effect of reducing the stock of investment funds available to banks relative to insurance companies have been found to importantly associate with increases in forward rates, which indicates a movement of the associated yield curve toward a more steeply rising or less steeply falling position, particularly in the earlier maturity regions.

## Appendix—Derivation of Zero-Coupon Equivalent Yields

*Dependent Variable:* Seasonally adjusted one year forward rate: ${}_{t+n}r_t$ where t is the date of observation, and $n$ = years after t that the forward rate applies.

The original data source for the dependent variable is monthly observations on the market price of U.S. Treasury bonds outstanding for the period 1964–1 through 1972–12, taken on the last trading day of each month. Yields are calculated from these prices, and are used to fit equation (A1), giving monthly yield curves on *U.S.* government bonds:[13]

$$\text{(A1)}\quad (1 + R_{t,n}) = C_0 \cdot C_1^{(1/n)} \cdot e^{C_2(n)} \cdot e^{C_3(x)} \cdot e^{\epsilon}.$$

where:

$R_{t,n}$ = the yield to maturity on an $n$-period bond, observed at time t,
$n$ = the time to maturity,[14]
$x$ = the coupon value,
$\epsilon$ = a white noise error term, and $C_0$, $C_1$, $C_2$, and $C_3$ are estimated coefficients.

The unadjusted forward rate at date t is derived from these yields by evaluating equation (1) for different values of $n$. The unadjusted forward rates were tested for a significant seasonal component using the Bureau of Census X-11 variant seasonal adjustment program.[15] Because a significant seasonal factor was found in every forward rate, all of the forward rate data were seasonally adjusted. The equivalent yield for a zero-coupon rate is found by setting $x$ equal to zero.

*Exogenous Variables*

$X(t)$ = Real *GNP*, expressed monthly by application of the spline function method as defined by the National Bureau of Economic Research (*NBER*) in their TROLL system library.[16] Source: U.S. Office of Business Economics (*OBE*).

[13] This yield curve model is developed and tested in the authors' paper (Echols and Elliott). The yield curve model is derived from the framework of term structure theory. It shows an average $R^2$ value of .85 over the 108 sets of bond price data used in the tests.

[14] The number of bonds outstanding at any date is a function of the date. Thus, for example, there are 35 observation values in January 1964 and 18 observed values in December 1972.

[15] The potential impact of a seasonal factor on interest rates is motivated by results obtained by Stanley Diller.

[16] The method uses Newton's divided-difference interpolation formula to obtain a system of equations for the second derivative of the spline function, evaluated at the knots. These equations are reduced so that they use only the second derivative and second derivative differences. Since the second derivative is linear and the start and end values are known to equal zero, the system can be solved.

$p(t)$ = *GNP* price deflator, expressed monthly by the *NBER* TROLL spline method. Source: *OBE*.

$B(t)$ = Total government purchases less total tax revenues, expressed monthly as above. Source: *OBE*.

$NE(t)$ = Net export balance, expressed monthly as above. Source: *OBE*.

$MS(t)$ = Monthly real money supply, defined as currency and demand deposits ($M_1$) divided by $p(t)$. Source for $M_1$: Federal Reserve System (*FRB*).

$R_{1,t}$ = Average market yield on 1-year U.S. Treasury Bills observed monthly. Source: *FRB*.

$P_{1,t}$ = Average monthly price on a zero-coupon equivalent 1-year bond, observed monthly, calculated according to the following:

$$P_{1,t} = (1 + R_{1,t})^{-1} \tag{A2}$$

$BH(t)$ = Monthly stock of bank funds invested in *U.S.* government securities at time t. Source: *FRB*.

$IH(t)$ = Monthly stock of insurance company funds invested in *U.S.* government securities at time t. Source: *FRB*.

$SG(t)$ = Monthly stock of outstanding 0–5 years to maturity *U.S.* government bonds at time t. Source: U.S. Treasury Department (*TB*).

$LG(t)$ = Monthly stock of outstanding 10-year-and-up years to maturity *U.S.* government bonds at time t. Source: *TB*.

## REFERENCES


L. C. Anderson and K. M. Carlson, "A Monetarist Model for Economic Stabilization," *Fed. Reserve Bank St. Louis Rev.*, Apr. 1970, *52*, 7–21.

D. Chambers and A. Charnes, "Inter-temporal Analysis and Optimization of Bank Portfolios," *Manage. Sci.*, Winter 1972, *7*, 393–410.

J. M. Culbertson, "The Term Structure of Interest Rates," *Quart. J. Econ.*, Nov. 1957, *71*, 485–517.

S. Diller, "Expectations in the Term Structure of Interest Rates," in J. Mincer, ed., *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*, New York 1969.

D. Durand, "Basic Yields of Corporate Bonds, 1900–1942," *Tech. Pap. 3*, Nat. Bur. Econ. Res., June 1942.

———, "A Quarterly Series of Corporate Basic Yields, 1952–57, and Some Attendant Reservations," *J. Finance*, Sept. 1958, *13*, 348–56.

M. E. Echols and J. W. Elliott, "A Quantitative Yield Curve Model for Estimating the Term Structure of Interest Rates," *J. Finance. Quant. Anal.*, Mar. 1976, *11*, forthcoming.

R. Eisner, "Research and Development and Other Determinants of Investment," *Amer. Econ. Rev. Proc.*, May 1969, *59*, 50–64.

J. W. Elliott, "Measuring Inflationary Expectations and the Real Rate of Interest: An Exploration of Macroeconomic Alternatives," mimeo., Univ. Wisconsin, Milwaukee, June 1975.

——— and M. E. Echols, "Liquidity Premiums, Interest Rate Risk, and Error Learning in the Term Structure," unpublished paper, Sch. Bus. Admin. Management Res. Center, Univ. Wisconsin, Milwaukee, Apr. 1975.

——— and ———, "Market Segmentation, Speculative Behavior, and the Term Structure of Interest Rates," *Rev. Econ. Statist.*, Feb. 1976, *58*.

M. Feldstein and O. Eckstein, "The Fundamental Determinants of the Interest Rate," *Rev. Econ. Statist.*, Nov. 1970, *52*, 363–75.

D. Fisher, "Expectations, The Term Structure of Interest Rates, and Recent British Experience," *Economica*, Aug. 1966, *33*, 319–29.

I. Fisher, *The Nature of Capital and Income*, New York 1906.

L. Fisher and R. L. Weil, "Coping with the Risk of Interest-Rate Fuctuations: Returns to Bondholders from Naive and Optimal Strategies," *J. Bus., Univ. Chicago*, Oct. 1971, *44*, 408–31.

W. E. Gibson, "Price Expectations Effects on Interest Rates," *J. Finance*, Mar. 1970, *25*, 19–34.

J. R. Hicks, *Value and Capital*, 2d ed., Oxford 1946.

J. Johnston, *Econometric Methods,* New York 1972.

N. Kaldor, "Speculation and Economic Stability," *Rev. Econ. Stud.,* Oct. 1939, *7,* 1–27.

R. A. Kessel, "The Cyclical Behavior of the Term Structure of Interest Rates," in J. M. Guttentag, ed., *Essays on Interest Rates, Vol. II,* New York 1971.

F. A. Lutz, "The Structure of Interest Rates," *Quart. J. Econ.,* Nov. 1940, *55,* 36–63.

J. H. McCulloch, "An Estimate of the Liquidity Premium," *J. Polit. Econ.,* Feb. 1975, *83,* 95–119.

B. G. Malkiel, "Expectations, Bond Prices, and the Term Structure of Interest Rates," *Quart. J. Econ.,* May 1962, *76,* 206–14.

———, "The Term Structure of Interest Rates," *Amer. Econ. Rev. Proc.,* May 1964, *54,* 532–43.

D. Meiselman, *The Term Structure of Interest Rates,* Princeton 1962.

J. B. Michaelson, "The Term Structure of Interest Rates and Holding Period Yields on Government Securities," *J. Finance,* Sept. 1965, *20,* 444–63.

F. Modigliani and R. Sutch, "Innovations in Interest Rate Policy," *Amer. Econ. Rev. Proc.,* May 1966, *56,* 178–97.

——— and ———, "Debt Management and the Term Structure of Interest Rates," *J. Polit. Econ.,* Aug. 1967, *75,* 569–89.

J. F. Muth, "Rational Expectations and the Theory of Price Movements," *Econometrica,* July 1961, *29,* 315–35.

C. R. Nelson, *The Term Structure of Interest Rates,* New York 1972.

R. Roll, *The Behavior of Interest Rates: The Application of the Efficient Market Model to U.S. Treasury Bills,* New York 1970.

J. Rutledge, *A Monetarist Model of Inflationary Expectations,* Lexington 1974.

T. J. Sargent, "Commodity Price Expectations and the Interest Rate," *Quart. J. Econ.,* Feb. 1968, *82,* 127–40.

———, "Rational Expectations and the Term Structure of Interest Rates," *J. Money, Credit, Banking,* Feb. 1972, *4,* 74–97.

———, "Interest Rates and Prices in the Long Run," *J. Money, Credit, Banking,* Feb. 1973, Part 2, *5,* 385–449.

R. H. Scott, "Liquidity and the Term Structure of Interest Rates," *Quart. J. Econ.,* Feb. 1965, *79,* 135–45.

W. T. Terrell and W. J. Frazer, Jr., "Interest Rates, Portfolio Behavior, and Marketable Government Securities," *J. Finance,* Mar. 1972, *27,* 1–36.

J. Tobin, "Liquidity Preference as Behavior Towards Risk," *Rev. Econ. Stud.,* Feb. 1953, *25,* 65–86.

J. Van Horne, "Interest-Rate Risk and the Term Structure of Interest Rates," *J. Polit. Econ.,* Aug. 1965, *73,* 344–51.

K. Wicksell, *Lectures on Political Economy,* Vol. II, New York 1935.

W. P. Yohe and D. S. Karnosky, "Interest Rates and Price Level Changes, 1952–1969," *Fed. Reserve Bank St. Louis Rev.,* Dec. 1969, *51,* 18–38.

Board of Governors of the Federal Reserve System, *Fed. Res. Bull. (FRB),* various issues, Washington.

U.S. Office of Business Economics *(OBE), Surv. Curr. Bus.,* various issues, Washington.

U.S. Treasury Department, *Treasury Bull. (TB),* various issues, Washington.

U.S. Bureau of the Census, "The X-11 Variant of the Census Method II Seasonal Adjustment Program," tech. pap. no. 15, Washington 1965.

# The Value of Human Life in the Demand for Safety

*By* BRYAN C. CONLEY*

The economics of safety is one of the interesting new areas to which economic theory is currently being applied.[1] Topics of analysis range from the estimation of costs of alternative safety programs to models of consumer choice incorporating safety aspects. From either an individual's or society's point of view, presumably there are safety precautions (or risks) worth taking depending upon costs (or benefits), the probabilities involved, wealth, risk aversion, and other factors.

The most critical safety decisions are those concerning the possibility of accidental loss of life. For a sufficiently small probability, there is a conceptual "value of human life" which, multiplied by the probability, will yield the maximum payment a person would make for the stated improvement in survival.

For government safety programs, the benefit per statistical life saved would be the average of the affected population's values of human life plus the value of any externalities. However, traditional benefit-cost analysis has relied upon the "human capital" approach, in which the benefit is the discounted earnings (reflecting productivity) of the affected population. Although more data are available on the human capital approach, the "willingness-to-pay" method is theoretically superior. So far no one has with success theoretically linked the two values, especially in terms of order. Indeed, some believe there is no necessary relationship.[2] Neither has a model been developed in which choices of specific commodities are involved and in which the probability of living is affected by such choices.

The purpose of this article is therefore twofold: 1) to extend the traditional model of individual maximization to include the effects of choices involving a changed probability of living; and 2) to determine the value of human life with reference to an individual's wealth and utility function characteristics.

In Section I, a deterministic model of individual optimization (except for time of death) is developed. The individual maximizes expected utility where the probability of death is under his control. After developing several forms of equilibrium conditions and interpreting them, I analyze the effect of relaxing two of the assumptions and summarize theoretical relationships published elsewhere. In Section II, the model is applied to the demand for safety, and implications are drawn for demand studies and for measurable utility.

The principal conclusion of this study is that for income above some undetermined but presumably low level, the value of life

* Assistant professor of business economics, Graduate School of Business, University of Southern California. Thanks are due Michael Intriligator, Finis Welch, and John Hambleton for helpful comments on an early draft, and an anonymous referee on a later draft. The Center for Public Economics, San Diego State University provided secretarial assistance. I am, of course, responsible for any errors.

[1] See Jack Carlson, Gary Fromm, Lester Lave, E. J. Mishan, Walter Oi, Thomas Schelling, and Joseph Spengler.

[2] Schelling states: "There is no reason to suppose that what a man would pay to eliminate some specific probability, *P*, of his own death is more than, less than, or equal to *P* times his discounted expected earnings. In fact there is no reason to suppose that a man's future earnings, discounted in any pertinent fashion, bear any particular relation to what he would pay to reduce some likelihood of his own death" (p. 149).

is greater than discounted earnings, and in early and middle adulthood, greater than discounted consumption.

## I. The Model

My aim is to separate the other uncertainties of living from those associated with risk of loss of life.[3] Thus it is assumed that all relevant variables and functional relationships are known, the only uncertainty being the actual time of death. Let a person be infinitely sensitive; have full information of all prices, probabilities, and his preferences; and be interacting in competitive markets with zero transaction costs.[4] Let him choose among life histories such as to maximize expected lifetime utility, the von Neumann-Morgenstern postulate. The objective function is of the form[5]

$$E(U) = \sum_{t} \delta(t)u(\cdot)p(\cdot) \tag{1}$$

whereby expected lifetime utility is separable into discounted single period utilities, $\delta(t)$ is the utility discount function, $u(\cdot)$ is a single period utility function, and $p(\cdot)$ is the probability of being alive.

The unit of analysis will be the decision or control variable $x^{it} \geq 0$, which is the value of activity $i$ in time period t. (Superscripts designate time and activity indices; subscripts indicate partial derivatives.) An activity will be called consumption if the consumer must expend cash outlays and production if it results in cash income.

I assume an infinite planning horizon.[6] Time 0 is the beginning of decision making at age $\tau$; and for time indices T and t, $0 \leq \mathrm{T} \leq \mathrm{t}$.

### *A. The Objective Function*

The utility discount function $\delta(\mathrm{t})$ shall take the form that R. H. Strotz proved was necessary for consistent multiperiod consumer behavior; thus

$$\delta(\mathrm{t}) \equiv \delta^{\mathrm{t}} \equiv \left(\frac{1}{1+s}\right)^{\mathrm{t}} \tag{2}$$

where $s$ is the subjective time discount rate of utility.

The single period time-invariant utility function shall be

$$u(\cdot) = u(x^{1t}, x^{2t}, \ldots, x^{nt}) = u(X^t) \tag{3}$$

where $X^t$ is the vector of activities in period t. Making the simplest of assumptions, I assume no externalities: none of the $x^{it}$ enter others' utility functions; neither do any of others' activities enter $u(\cdot)$. Thus there is no family that cares about a person's life and no desire to leave an estate is allowed. Below I shall relax this assumption.

The utility function $u$ is assumed to be strictly concave, i.e., $u'' < 0$; we will not require either $u' > 0$ or even $u > 0$ for all values of every activity. Assume $u(0, 0, \ldots, 0) \equiv u^* = 0$; that is, the utility of not living is zero.[7] Likewise it is assumed there are some $X^t > 0$ for which $u(X^t) = 0$.[8] Thus

[3] This approach is also adopted by Menahem Yaari.

[4] Oi's model also depends on these assumptions.

[5] Expressing lifetime expected utility in continuous form appears to be impossible due to difficulties in expressing the $p$ function. In a related paper, Morton Kamien and Nancy Schwartz construct a similar model for optimal maintenance of a machine. They assume that maintenance expenditures affect only the immediate probability of machine failure, which is more restrictive than my initial behavioral assumption contained in equation (4) below. Subsequently in equation (9) attention is restricted to activities which immediately affect probability of survival. Had I so restricted the assumptions initially, the model could have been developed in an optimal control framework with identical results.

[6] For convenience, no upper bound on a lifespan is assumed. No arbitrary upper bound on an individual's planning horizon can be postulated, as there is no future time for which $p(\cdot) = 0$ by necessity. Although an infinite life has not yet been experienced, the history of mortality cannot logically negate its possibility. Also, future medical advances may make an infinite life biologically possible.

[7] This assumption is basically atheistic in orientation. Below, I shall explore the implications of altering this assumption.

[8] Some utility functions such as $u = \log x$ and $u = C_0 - x^\gamma$ have this property; this level of consumption may be considered a subsistence bundle of goods, although

$u$ is determinable up to a multiplicative positive constant.

The probability-of-living function $p(\cdot)$ through period t, conditional on being alive at the beginning of period T, is[9]

$$(4) \quad p^{Tt}(T, t, x^{1T}, x^{2T}, \ldots, x^{nT}, x^{1T+1}, x^{2T+1}, \ldots, x^{nT+1}, \ldots, x^{1t}, \ldots, x^{nt}) \equiv p^{Tt}(T, t, {}_{Tt}X)$$

where ${}_{Tt}X$ is the vector of $X^k$ from time T to t; $p^{Tt}(\cdot)$ is an objective conditional probability function known to all participants. Death of a person at the end of a given period may be thought of as due to the "state of nature" reacting with his activities, where the state of nature is not perfectly forecast. Note that $p^{0t} = p^{0T}p^{Tt}$ and $0 \leq p^{Tt} < 1$ for all T and t.

In summary, the total expected lifetime utility function to be maximized at any period T is

$$(5) \quad U^T = E(U)^T = \sum_{t=T}^{\infty} \left(\frac{1}{1+s}\right)^{t-T} u(X^t)p^{Tt}(T, t, {}_{Tt}X)$$

### B. Constraints

The monetary constraint requires that expected discounted lifetime consumption $C$ equal expected discounted lifetime labor income $Y$ plus wealth $W$ at T: $C^T = Y^T + W^T$, or

$$(6) \quad \sum_{t=T}^{\infty} \left(\frac{1}{1+r}\right)^{t-T} \sum_{i=1}^{n} \$^{it}x^{it}p^{Tt}(T, t, {}_{Tt}X) - W^T = 0$$

where $r$ is the market interest rate (lending and borrowing rate, assumed constant indefinitely), and $\$^{it}$ is the monetary value associated with each unit of $x^{it}$ such that $\$^{it} > 0$ when $x^{it}$ is a consumptive activity and $\$^{it} < 0$ when $x^{it}$ is productive.[10]

In full-information, competitive, costless markets such a constraint is feasible, since one can contract with an insurance company in period T to engage in activities ${}_{T\infty}X$, turning over all wealth and future labor income and receiving an allowance for contractual consumption expenditures. Therefore if (6) holds, the expected value at T of the company's receipts will equal the expected value at T of its expenditures, a contract it would write given the foregoing assumptions.

The above contract is similar to the "commodity-options" of Kenneth Arrow and the "actuarial notes" of Menahem Yaari. The purpose of the contract is not to provide for others' well-being upon one's death, but rather to expand one's expected lifetime utility.[11] In effect when one's past consumption exceeds one's past income, the insurance upon death pays off one's debt. It is not "full insurance"[12] because one is not indifferent as to whether or not the "event" occurs.

Such a contract is optimal (see Arrow), even though the probabilities are affected by one's behavior, because the company is aware of such behavior and adjusts its contractual allowances accordingly. The models of Isaac Ehrlich and Gary Becker, Walter Oi and Lester Lave incorporate the assumption of full information which leads to optimal insurance coverage.[13] Thus, as

life is possible at a lower level of consumption. If expected lifetime utility $U^T$ is negative, that is, life is less desirable than being dead, we assume a person commits suicide.

[9] Equation (4) may be considered a production function linking the probability of survival to various inputs. However it differs from most production functions in that the functional relationship is behaviorally unique to each individual.

[10] This constraint was used by Yaari and Nils Hakansson.

[11] Given that an individual wishes to leave no estate, a comparable annuity program is the optimal method of providing for retirement. Even if such a contract is not entered into, all of the subsequent conclusions are identical if a person can borrow against future earnings secured by term insurance.

[12] Isaac Ehrlich and Gary Becker state: "Full insurance can be identified with full coverage of potential losses" (p. 627n).

[13] Lave states: "A liability insurer would be motivated

Ehrlich and Becker point out, the "moral hazard" problem is eliminated.

In complete form the optimization equation (at T=0 and assuming $W^0=0$) becomes

$$(7)\quad \underset{[x^{it}]}{\mathrm{Max}}\ U^0=E(U)^0 = \sum_{t=0}^{\infty}\left(\frac{1}{1+s}\right)^t u(X^t)p^{0t}(0,\text{t},{}_{0t}X) - \lambda^0\left[\sum_{i=1}^{\infty}\left(\frac{1}{1+r}\right)^t\sum_{t=0}^{n}\$^{it}x^{it}p^{0t}(0,\text{t},{}_{0t}X)\right]$$

*C. Optimization*

The optimization of (7) for any activity $x^{iT}$ leads to the first-order condition (for an interior maximum):

$$(8)\quad \left(\frac{1}{1+s}\right)^{T}u_{iT}p^{0T}+\sum_{t=T}^{\infty}\left(\frac{1}{1+s}\right)^{t}u^{t}p_{iT}^{0t} -\lambda^{0}\left[\left(\frac{1}{1+r}\right)^{T}\$^{iT}p^{0T} + \sum_{t=T}^{\infty}\left(\frac{1}{1+r}\right)^{t}\sum_{k=1}^{n}\$^{kt}x^{kt}p_{iT}^{0t}\right]=0$$

Concavity assures the second-order conditions will be met.

This paper is concerned with those safety decisions which have the characteristic that they affect the current probability of living, but, if one survives, do not affect later conditional probabilities of survival.[14] Thus at the margin the change in the probability of survival through t of an activity undertaken in T is the probability of survival from 0 to $\dot{T}$ times the changed probability of survival through T times the conditional probability of survival from T through t. Thus

$$(9)\qquad p_{iT}^{0t} = p^{0T}\cdot p_{iT}^{TT}\cdot p^{Tt}$$

Expected lifetime net consumption at time T is

$$(10)\quad \sum_{t=T}^{\infty}\left(\frac{1}{1+r}\right)^{t-T}\sum_{i=1}^{n}\$^{it}x^{it}p^{Tt} = C^T - Y^T$$

Let $V^T$ equal net labor income; $V^T=Y^T-C^T$. Therefore $V^T$ is the expected discounted net receipts of the insurance company from its contract. Although the actuarial value of the contract is zero at time 0, if a person consumes more than he earns in early adulthood (a likely result of the "hump savings" type of optimal consumption), then the insurance company has paid out more than it has received, and therefore expected future net receipts must be positive. The value $V^T$ represents the net loss to others (here the insurance company) from death at T.

Next the Fisherian interperiod equilibrium equation is introduced,

$$(11)\qquad \lambda^T = \lambda^0\left(\frac{1+s}{1+r}\right)^T$$

where $\lambda^T=\partial U^T/\partial W^T$, the marginal utility of wealth at T. Yaari has shown (11) holds for decisions made under uncertainty when insurance is available.

Substituting (5), (9), (10), and (11) into (8) and dividing by $p^{0T}$ yields

$$(12)\quad u_{iT} = -\,p_{iT}^{TT}U^T + \lambda^T[\$^{iT} - p_{iT}^{TT}V^T]$$

The interpretation of (12) for $u_{iT}>0$, $p_{iT}^{TT}<0$, and $\$^{iT}>0$ is as follows: in equilibrium, the marginal utility of activity $i$ must equal the sum of the expected loss of lifetime utility $-p_{iT}^{TT}U^T$, plus the marginal utility of the sum of the monetary cost of

to determine the *total* cost of each safety-related act; it would be motivated to gear its premium schedule to the total cost of each safety-related act and so internalize the entire cost of the act. With liability laws, the reduction in my insurance premium stemming from fixing my brakes would reflect the cost to the people I was likely to strike as well as the cost to my property. This premium would be of the correct size to motivate an optimal level of safety" (p. 520).

[14] Many activities such as medical checkups, smoking, nonfatal injuries, and body care affect future conditional probabilities of living. The equilibrium conditions for such activities will not be explored further.

activity $i$ and the change in expected net labor income $\lambda^{T}(\$^{iT}-p_{iT}^{TT}V^{T})$.

Rearranged, (12) becomes

$$(13)\qquad \lambda^{T}=\frac{u_{iT}+p_{iT}^{TT}U^{T}}{\$^{iT}-p_{iT}^{TT}V^{T}}$$

If $p_{iT}^{TT}=0$, (13) becomes $\lambda^{T}=u_{iT}/\$^{iT}$, the familiar equilibrium condition.

Equation (13) may be further manipulated to generate a more interesting result. Recognizing that marginal utility of lifetime consumption equals the marginal utility of lifetime wealth,[15] even though not all of an exogenous increase in wealth would be used for future consumption, we have $\lambda^{T}\equiv\partial U^{T}/\partial W^{T}=\partial U^{T}/\partial C^{T}$. Dividing (12) by $\lambda^{T}$ and rearranging yields[16]

$$(14)\quad \$^{iT}-p_{iT}^{TT}\left(\frac{U^{T}}{\partial U^{T}/\partial W^{T}}+V^{T}\right)=u_{iT}/\lambda^{T}$$

The term $U^{T}/(\partial U^{T}/\partial W^{T})$ may be manipulated to yield

$$(15)\quad \frac{U^{T}}{\partial U^{T}/\partial W^{T}}\equiv\frac{C^{T}}{\dfrac{\partial U^{T}}{\partial C^{T}}\dfrac{C^{T}}{U^{T}}}\equiv\frac{C^{T}}{\alpha^{T}}\equiv L^{T}$$

where we define $\alpha^{T}\equiv(\partial U^{T}/\partial C^{T})/(C^{T}/U^{T})$, which is the lifetime consumption elasticity of lifetime utility, and $L^{T}\equiv C^{T}/\alpha^{T}$. Substituting (15) into (14) and rearranging yields this alternative formulation for the equilibrium condition:

$$(16)\qquad \lambda^{T}=\frac{u_{iT}}{\$^{iT}-p_{iT}^{TT}(L^{T}+V^{T})}$$

which is similar to (13) except that the term $p_{iT}^{TT}U^{T}$ in the numerator of (13) has been incorporated in the denominator of (16) as $-p_{iT}^{TT}L^{T}$.

### *D. Interpretation*

Before interpreting (16), I shall interpret the term $H^{T}\equiv L^{T}+V^{T}$ which appears in all equilibrium equations at time T. Here it is claimed that $H^{T}$ represents the *value of life saving* to a person appropriate to valuing safety issues.

This claim may be verified by multiplying (16) by $\$^{iT}-p_{iT}^{TT}H^{T}$ to derive

$$(17)\qquad \lambda^{T}(\$^{iT}-p_{iT}^{TT}H^{T})=u_{iT}$$

For a pure safety activity, $u_{iT}=0$ and therefore

$$(18)\qquad \$^{iT}=p_{iT}^{TT}H^{T}$$

That is, the amount the individual will pay for the pure safety activity is the changed probability of living times $H^{T}$; therefore $H^{T}$ is the value of an individual's life used in evaluation safety decisions, as discussed in the introduction.

The term $V^{T}$ represents others' interest

---

[15] In the maximization of

$$\text{(a)}\quad \bar{U}(C,L)=U(C,L)-\lambda(C-(\bar{L}-L)w-W)$$

where $C$ is consumption, $L$ hours of leisure (time not working), $W$ wealth, $\bar{L}$ maximum hours available per time period (e.g., 168 hours per week), and $w$ the wage rate, one first-order condition is

$$\text{(b)}\quad \frac{\partial\bar{U}}{\partial C}=\frac{\partial U}{\partial C}-\lambda=0$$

[16] Equation (14) becomes similar to the equilibrium condition of Fromm when $V^{T}=0$, $u^{iT}=0$, the definitions $\$^{iT}\equiv\partial W^{T}/\partial x^{iT}$ and $p_{iT}^{TT}\equiv\partial p^{TT}/\partial x^{iT}$ are substituted, and (14) becomes:

$$\text{(a)}\quad \frac{\partial W^{T}}{\partial p^{TT}}=\frac{U^{T}}{\partial U^{T}/\partial W^{T}}$$

(One must set the desire for an estate and disutility of the probability of death in Fromm's model equal to zero.) Equation (a) also follows directly from maximizing $V=pU$:

$$\text{(b)}\quad \frac{\partial V}{\partial p}=U+p\frac{\partial U}{\partial W}\frac{\partial W}{\partial p}=0$$

Thus

$$\text{(c)}\quad -\frac{\partial W}{\partial p}=\frac{U}{p(\partial U/\partial W)}=\frac{U}{\partial U/\partial W}$$

where $p\simeq1$. This is the original form in which the problem was solved, including this additional manipulation to be discussed in the text:

$$\text{(d)}\quad -\frac{\partial W}{\partial p}=\frac{U}{\partial U/\partial W}=\frac{CU}{CU'}=\frac{C}{\alpha}$$

$$=\text{value of human life.}$$

in preserving life, while $L^T$ which we label *the value of human life* is the individual's own interest in his safety decision. In this full-information model, however, others' interest in safety is internalized into the individual's decision-making process.

We may now interpret (16). For activities affecting safety for which $u^{iT}>0$, $\$^{iT}>0$, $p_{iT}^{TT}<0$, $L^T>0$, and $V^T>0$, the total cost (or full price) of the activity in equilibrium is comprised of three parts: the monetary cost $\$^{iT}$, the expected loss of the value of human life $-p_{iT}^{TT}L^T$, and the expected loss to others (the insurance company) of net labor income, $-p_{iT}^{TT}V^T$. Of course, $p_{iT}^{TT}$, $V^T$, $\$^{iT}$, and $u^{iT}$ may take on positive and negative values in the real number system, but (16) must hold.

In particular, what values might $L^T$ take on? Since $L^T=U^T/U'^T$, $L^T>0$ since $U^T>0$ by assumption. (See fn. 8.) In general, $L^T>C^T$ holds whenever $\alpha^T<1$, a condition that we expect generally holds for most values of $C^T$ but that as shown below does not necessarily hold for any value of $C^T$. Figure 1a presents the typical case in which a strictly concave lifetime utility function with $U'>0$ will have three regions. Here let $U_B=U^*=0$, where $U^*$ is utility associated with being dead. In the first region $U^T<0$, and thus $\alpha^T<0$ (see Figure 1b). In the second, $U^T>0$ but $\alpha^T>1$ and thus $L^T<C^T$ (see Figure 1c). In the third, which may be called the general case, $0<\alpha^T<1$, and $L^T>C^T$; that is, for a value of expected lifetime consumption above some critical value where $\alpha^T=1$, the value of human life exceeds expected lifetime consumption. This critical value is presumably at a low level of income. However, it is not ever necessary that $\alpha^T<1$ for any $C^T$. In Figure 2, $\alpha^T>1$ for all $C^T>C^0$. In all cases $L^T$ increases more than $C^T$ if income increases since

$$(19)\quad \frac{\partial L^T}{\partial C^T}=\frac{U'U'-UU''}{U'^2}=1+L^TR^T>1$$

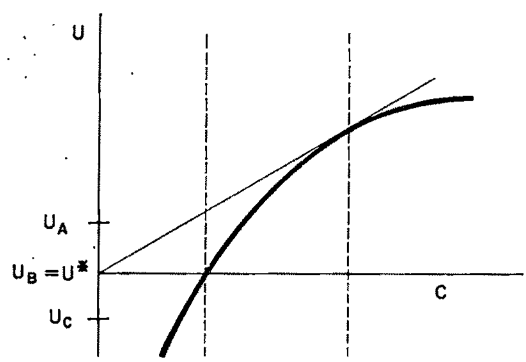


FIGURE 1a

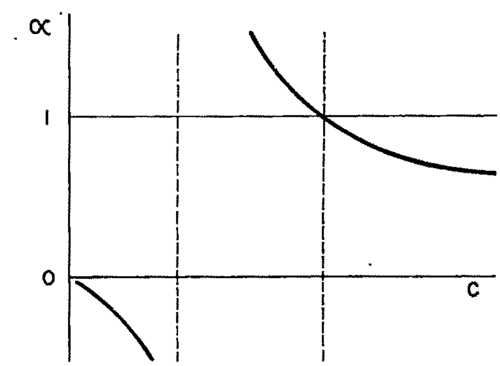


FIGURE 1b

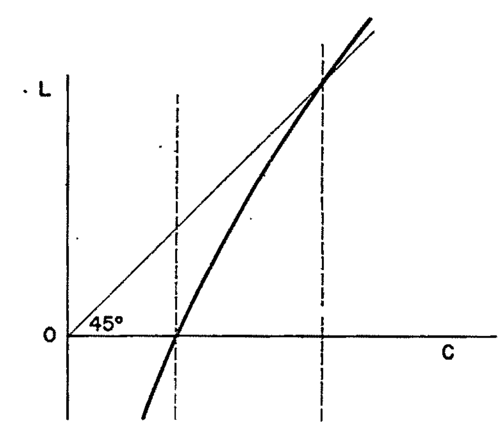


FIGURE 1c

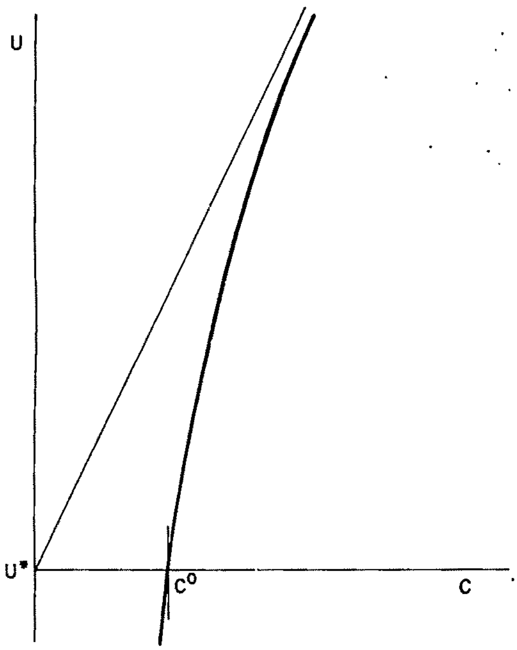



FIGURE 2

where $R^{\mathrm{T}} = -U''/U' > 0$ is the Pratt-Arrow absolute risk-aversion index.

The value of life saving $H^{\mathrm{T}}$ will generally have values close to $L^{\mathrm{T}}$, since their difference $V^{\mathrm{T}}$ will generally be of smaller magnitude than that of $L^{\mathrm{T}}$, except near or in retirement.[17]

The major theoretical conclusion may now be stated. Since the value of life saving $H = L + V = C/\alpha + (Y - C) = C(1/\alpha - 1) + Y$ in the general case where $\alpha < 1$, $H > Y$; that is, the value of life saving is greater than discounted lifetime labor income. Furthermore, in the early and middle years of adulthood when $V > 0$ (due to early dissaving), then the value of life saving in the general case will be greater than discounted lifetime consumption.

Equation (16), in conjunction with (6) and (11), determines lifetime equilibrium conditions. Of the eighty-one combinations of sign values which the vector of variables $[u_{i\mathrm{T}}, \$^{i\mathrm{T}}, p_{i\mathrm{T}}^{\mathrm{TT}}, V^{\mathrm{T}}] \equiv S$ may take on (all may be $+$, 0, or $-$), almost all are of compatible signs. Of these, only a few are of special interest.

When $S = [\pm, -, -, \pm]$, a person is undertaking a risky job, and is compensated with higher pay for the increase in risk.

When $S = [-, +, +, \pm]$, a person is paying for what might be called a "preventive consumption activity," for which his marginal utility is negative; but the activity, such as taking an emetic after an accidental overdose of a drug, considerably increases his probability of life.

Finally, when $S = [0, +, +, \pm]$, a person is spending for a "pure safety activity," in which the only value of the activity is an improved probability of living. This type of expenditure is often relevant for public safety expenditures.

In my earlier paper, I have shown the effects of relaxing some of the assumptions used above. Here I will explore only two of them.

If the utility level associated with death (or of not being alive) is not arbitrarily set at zero, then

$$u(0, 0, \ldots, 0) \equiv u^* \neq 0 \tag{20}$$

In Figure 1a, $U_B = U^*$ and either $U_A$ or $U_C$ instead of $U_B$ may equal zero.

Optimizing

$$V = pU + (1 - p)U^* \tag{21}$$

yields

$$\frac{\partial C}{\partial p} = \frac{U - U^*}{\partial U/\partial C} = L \tag{22}$$

where $p \simeq 1$. Then the numerator of the value-of-life function is the *difference* in utility levels between living and not living. Those who anticipate a pleasant hereafter such that $U$ and $U^*$ do not differ by much therefore will not place a high value on safety expenditures.

[17] For a fuller discussion, see my unpublished paper.

Finally we can incorporate the implications of a person's desire to leave an estate. For convenience, let activity 1 be dollar consumption by others, financed by the person in question. Then $\$^{1t}=1$ and $p_{1t}^{tt}=0$. Thus $u(x, 0, 0, \ldots, 0)\equiv u^*(x^{1t})$ represents one's welfare in death based upon others' consumption. If $X^{1t}$ is the discounted value of such consumption, then $U^*(X^{1t})$ would represent discounted utility of a person who died in period t.

The inclusion of a desire for others' consumption has two implications. First, as more of potential income is allocated for the consumption of others, one's own consumption is necessarily reduced and most likely $L$ will be less (since we cannot arbitrarily restrict the value of $\alpha$, the conclusion must be tentative).

Secondly, a person most likely will increase his lifetime welfare by allocating potential income to others *whether he lives or not.* It would not be unreasonable to assume that whether or not one is alive, one's utility would be increased through others' consumption by about the same degree. Therefore the relationship between $U$ and $U^*$ and the value of $U'$ should be relatively unchanged. If lifetime income and therefore consumption should increase (here at time 0, $C^0=Y^0$), $U$ and $U^*$ both increase, but the former at a faster rate, and $\partial U/\partial C$ declines so that $\partial L/\partial C>1$ and, in the general case, above some level of consumption, $L>C$.

This relationship is shown graphically in Figure 3. The total utility curve at period T is segmented into two parts: 1) utility $\hat{U}$, if those monies planned for others' consumption while alive were not available, and 2) the difference $U-\hat{U}$, which is the increase in utility due to such transfers. At income levels $Y_2$ and $Y_3$, the increase in utility $U_2-\hat{U}_2$ and $U_3-\hat{U}_3$, respectively, due to such transfers is shown to be approximately equal to the utility derived from leaving an estate, $U_2^*$ and $U_3^*$, respectively.

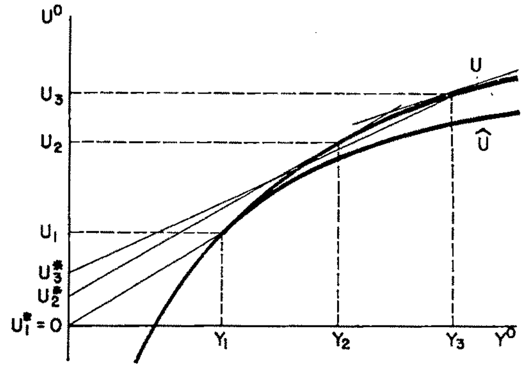



FIGURE 3

At income level $Y_1$, a person allocates nothing for others either while living or in an estate so the graphical presentation is the same as in Figure 1a.

At income levels $Y_2$ and $Y_3$, however, some monies are available for others whether a person lives or not. At $Y_2$, since $Y_2\cdot\partial U/\partial Y=U_2-U_2^*$, the value of human life equals the present value of income, that is, $L=Y=C$. At $Y_3$, the slope of the ray between $U_3$ and $U_3^*$ is greater than that of the tangent to $U$ at $Y_3$. Thus $L_3>C_3$. In general the inclusion of a desire for an estate most likely would not materially change the relationships developed when no desire for any transfers was assumed.

In my earlier paper I extended the basic model to several related aspects. The compensated demand elasticity for safety was found to be unitary, and the uncompensated demand elasticity is bounded by the interval $(-1/2, 0)$. I developed the supply curve[18] and found that safety is a normal noninferior good. Finally, if $\alpha<1$ and if the time $t_c$, when the discounted value of lifetime consumption reaches a maximum, precedes the time $t_y$, when the

[18] Safety is a joint output of household production, and thus the supply curve for safety reflects the increased expenditures necessary to secure greater safety while keeping current utility levels constant.

discounted value of lifetime income reaches a maximum (both for a given income stream), then $L$ and $H$ reach maxima before the discounted value of income reaches a peak, or

$$0 \leq t_L \leq t_H < t_y \tag{23}$$

Therefore the most valuable lives to be saved for a given lifetime income stream are those in their twenty's, early thirty's or possibly younger.

In this section we have determined that all individuals have a value of life $L^T$ which, added to the present value of net income $V^T$, theoretically is used in evaluating activities affecting their probability of living. In general $L^T$ is a multiple of the present value of consumption by the inverse of $\alpha^T < 1$. A heuristic interpretation of $L^T$ is as follows: Consider a person whose $\alpha^T = .2$ and who is considering a prospect involving a 1 percent increase in the probability of living through a period and involving a monetary cost. If he maximizes expected utility and is indifferent between his present circumstances and the prospect, then he must expect 1 percent less welfare due to the monetary cost if he survives the period (for which there is a 1 percent greater chance). Since $\alpha^T = .2$, his lifetime expected consumption upon survival must drop by 5 percent for him to be indifferent to the prospect.

## II. Applications

In this section I consider three broad areas where the model may be applied. The first obvious application is to the demand for safety. Second, applications to demand for other commodities, such as transportation, are explored. Finally implications for measurable utility are drawn.

### *A. Demand for Safety*

The social benefit of a pure safety activity (where $S = [0, +, +, \pm]$) is $p_{iT}^{TT}(L^T + V^T + E^T)$, where $E^T$ is the value of net nonmonetary externalities. At best private benefits equal $p_{iT}^{TT}(L^T + V^T)$. Therefore the private demand for safety will differ from the socially optimal amount depending upon the value of $E^T$.

If government undertakes a pure safety activity, say the $k$th, then the total of social benefits will be

$$B_k^T = \sum_{j=1}^{m} {}^j p_k^{TT} ({}^j L^T + {}^j V^T + {}^j E^T) \tag{24}$$

where $B_k^T$ is social benefits and the index $j$ refers to persons in the affected population. If some individuals do not have an insurance contract as developed in this paper, then ${}^j V^T$ equals the present value of the estate and transfers (including support of family), i.e., net income plus net worth, that one was planning to transfer in the future.

The model suggests that an individual is symmetrical as regards increased or decreased risks. If he accepts $R$ for an increased probability of death $p'$ (where $p' \simeq 0$), then he will pay $R$ for a reduced probability of death $p'$. If individuals are observed to be demanding different amounts, there are other aspects such as transaction costs or anxiety which must be incorporated in the model.

The above analysis holds only for $p_{iT}^{TT}$ sufficiently small such that the marginal conditions are not significantly affected. Equation (16) applies in the small but not in the large. If a person wishes to pay \$5 for a 1/100,000 improvement in the probability of living, it does not follow that he or she will pay \$5,000 for a 1/100 improvement in living, or be willing to receive \$250,000 for taking a one-half chance of losing his or her life. Indeed, in the latter example, many persons could quite rationally not accept any amount of money for the chance.[19] Thus scaling is not linear,

[19] If such is the case, and if people are expected utility maximizers, it would prove that the utility function $U$ is bounded from above.

as Thomas Schelling has proposed.[20]

Finally, the above model does confirm the analysis of James Buchanan and Oi that the doctrine of consumer liability (when the consumer is fully informed) would lead to optimal results, since the efficient full price facing individuals would vary according to their preferences and wealth.

Empirical evidence of the value of human life in various circumstances is sketchy. Schelling conjectures that a statistical life saved may be worth two or three lifetime incomes. Jack Carlson reports values from $200,000 to a million dollars per life saved in *U.S.* defense expenditures and pilot compensation for risky flying. Richard Thaler and Sherwin Rosen estimated the risk premium that 907 individuals demanded for dangerous jobs; they found a resulting value of human life of around $200,000, which is several times lifetime earnings. Thus what evidence is available tends to confirm the major theoretical finding of this paper, that the value of life is greater than discounted earnings.

### *B. Demand Studies*

Since a change in safety is one factor in the cost (or benefit) of an activity, such a change will have predictable effects upon demand. Specifically, if $\eta^{i\mathrm{T}}$ is full price elasticity of a commodity ($u_{i\mathrm{T}}>0$), where the full price $= s^{i\mathrm{T}} =$ denominator of (16), then the "probability-of-living" elasticity of demand for good $i$, $\epsilon^{i\mathrm{T}}$, is

$$(25)\quad \epsilon^{i\mathrm{T}} = \frac{\partial x^{i\mathrm{T}}}{\partial p_{i\mathrm{T}}^{\mathrm{TT}}}\frac{p_{i\mathrm{T}}^{\mathrm{TT}}}{x^{i\mathrm{T}}} = \frac{\partial s^{i\mathrm{T}}}{\partial p_{i\mathrm{T}}^{\mathrm{TT}}}\frac{\partial x^{i\mathrm{T}}}{\partial s^{i\mathrm{T}}}\frac{s^{i\mathrm{T}}}{x^{i\mathrm{T}}}\frac{p_{i\mathrm{T}}^{\mathrm{TT}}}{s^{i\mathrm{T}}}$$

$$= \frac{p_{i\mathrm{T}}^{\mathrm{TT}}\cdot H^{\mathrm{T}}}{s^{i\mathrm{T}}}\cdot \eta^{i\mathrm{T}}$$

and therefore $\epsilon^{i\mathrm{T}}$ equals $\eta^{i\mathrm{T}}$ times the percentage of the full price of the $i$th commodity attributable to the value of life saving, namely $p_{i\mathrm{T}}^{\mathrm{TT}}H^{\mathrm{T}}/s^{i\mathrm{T}}$. This formula is identical to that of the price elasticity of demand for an input where the elasticity of substitution between the input and all other inputs in a production function is zero. Where $u_{i\mathrm{T}}>0$ and $p_{i\mathrm{T}}^{\mathrm{TT}}>0$, an improvement in the safety of a product *must* increase the demand for it, contrary to Oi's result that for some goods an improvement in quality may decrease market demand.[21]

Including a safety variable can improve the econometric estimation of demand for goods, such as for many sectors of transportation involving a safety aspect. The improvement in safety may predict part of the unexplained increased in demand for, as an example, commercial air transportation.

### *C. Measurement of Utility*

Previous attempts at measuring utility have been unsuccessful. However, if data relevant to safety decisions were available, they would help us measure the shape of such a function, if the assumptions of this model are appropriate. If independent measurements of $p_{i\mathrm{T}}^{\mathrm{TT}}$, $\$^{i\mathrm{T}}$, $C^{\mathrm{T}}$, and $V^{\mathrm{T}}$ can be made for pure safety activities, $L^{\mathrm{T}}$ and therefore $\alpha$ may be estimated for a wide range of incomes.

Suppose that empirical data generated average values of human life at a given age

[20] See p. 152–54. Also see Martin Bailey, p. 164, regarding scaling.

[21] Oi's model was for the demand for market goods, a percentage of which were defective and inflicted a known money cost. Market demand would have positive elasticity to changes in risk when (p. 11)

$$\text{(a)}\qquad -\eta_p < \frac{1}{1+\dfrac{W}{P}}$$

where $\eta_p$ is the market price elasticity for risky activities, $P$ is market price, and $W$ is net damage cost per defective unit. In our model $W=H^{\mathrm{T}}$, which in the general case makes the right-hand side of (a) imperceptably greater than zero. Therefore demand for any good is negatively related to its risk.

TABLE 1—PRESENT VALUE OF CONSUMPTION AND VALUE OF HUMAN LIFE (HYPOTHETICAL)

| Consumption | \$30,000 | \$70,000 | \$100,000 | \$200,000 |
|---|---|---|---|---|
| $L$ | \$10,000 | \$70,000 | \$150,000 | \$400,000 |
| $\alpha$ | 3 | 1 | 2/3 | 1/2 |

as hypothesized in Table 1. If people are expected utility maximizers, then the general shape of the total utility curve may be derived from such data.

Other general applications of the model are possible, such as to optimal population theory, where the objective may be to maximize discounted utilities. The model in general therefore continues the modification of classical consumer theory to broaden and modify the simplistic Hicks-Allen representation of consumer optimization.

## REFERENCES

**K. J. Arrow**, "Economic Welfare and the Allocation of Resources for Invention," in *The Rate and Direction of Inventive Activity: Economic and Social Factors*, Universities-Nat. Bur. Econ. Res. conference series, Princeton 1962, 609–25.

**M. J. Bailey**, "Comments [on Schelling]," in S. B. Chase, Jr., ed., *Problems in Public Expenditure Analysis*, Washington 1968, 162–66.

**J. M. Buchanan**, "In Defense of Caveat Emptor," *U. Chicago Law Rev.*, Fall 1970, *38*, 64–73.

**J. W. Carlson**, "Valuation of Life Saving," unpublished doctoral dissertation, Harvard Univ. 1963.

**B. C. Conley**, "The Value of Human Life in the Demand for Safety," unpublished paper, Center Publ. Econ., San Diego State Univ., Nov. 15, 1973.

**I. Ehrlich and G. S. Becker**, "Market Insurance, Self-Insurance, and Self-Protection," *J. Polit. Econ.*, July/Aug. 1972, *80*, 623–48.

**G. Fromm**, "Comments [on Schelling]," in S. B. Chase, Jr., ed., *Problems in Public Expenditure Analysis*, Washington 1968, 166–76.

**N. H. Hakansson**, "Optimal Investment and Consumption Strategies Under Risk, an Uncertain Lifetime, and Insurance," *Int. Econ. Rev.*, Oct. 1969, *10*, 443–66.

**M. I. Kamien and N. L. Schwartz**, "Optimal Maintenance and Sale Age for a Machine Subject to Failure," *Manage. Sci.*, Apr. 1971, *17*, B495–B504.

**L. B. Lave**, "Safety in Transportation: The Role of Government," *Law Contemp. Probl.*, Summer 1968, *33*, 512–35.

**E. J. Mishan**, "Evaluation of Life and Limb: A Theoretical Approach," *J. Polit. Econ.*, July/Aug. 1971, *79*, 687–705.

**W. Y. Oi**, "The Economics of Product Safety," *Bell J. Econ.*, Spring 1973, *4*, 3–28.

**T. C. Schelling**, "The Life You Save May Be Your Own," in S. B. Chase, Jr., ed., *Problems in Public Expenditure Analysis*, Washington 1968, 127–62.

**J. J. Spengler**, "The Economics of Safety," *Law Contemp. Probl.*, Summer 1968, *33*, 619–38.

**R. H. Strotz**, "Myopia and Inconsistency in Dynamic Utility Maximization," *Rev. Econ. Stud.*, no. 3, 1956, *23*, 165–180.

**R. Thaler and S. Rosen**, "Estimating the Value of a Life," paper presented at The Workshop in Human Resource Economics, UCLA, Feb. 19, 1974.

**M. E. Yaari**, "Uncertain Lifetime, Life Insurance, and the Theory of the Consumer," *Rev. Econ. Stud.*, Apr. 1965, *32*, 137–50.

# The Demand for High-Powered Money

*By* JAMES R. LOTHIAN*

The resurgence of the quantity theory in the past two decades has brought with it a renewal of an often-debated question—how to define money. Though the answers that have been given vary widely, most economists, with little hesitation, would probably settle on a monetary total similar to $M_1$ or $M_2$ in the United States as "the" definition of money.

What this paper shows, however, is that the choice is not nearly so simple. Empirical analysis is often circumscribed by lack of data on many of the variables theory suggests can affect the demand for money, such as the interest rates paid on various categories of deposits and the quality or "moneyness" of those deposits. Under some circumstances the omission of these variables can have a serious effect on the stability of conventional deposit-inclusive definitions of money, causing them to be highly imperfect indicators of the effects of money on the overall economy.

The solution that I propose in these situations is to return to a narrower definition of money, high-powered money alone. The rationale is that since high-powered money is of relatively constant quality over time and space, such specification errors are likely to be less important for high-powered money than for deposit-inclusive totals. The demand for high-powered money should be more stable than the demand for other monetary aggregates.

Since differences in deposit quality and deposit interest rates are likely to be particularly great across countries, I test this proposition by analyzing the demand functions for high-powered money and for various other monetary assets estimated across an international time-section sample spanning forty countries in the postwar period. On the whole the results support my hypothesis. I find that high-powered money is unambiguously the most stable total across the countries in my sample, even when it is judged on the basis of the constant velocity model and the other totals are judged on the basis of a more sophisticated model of money demand. Moreover, I find evidence of several sorts that the factors which a priori could be expected to produce this result have actually been operable.

## I. Theoretical Considerations

### A. *Money Demand Functions*

To be able to make a judgment about which monetary total is most stable in demand, and hence which is the most useful definition of money, we first have to decide which variables we will assume to be the most important determinants of demand and which we will assume to have only a negligible impact and will ignore in the statistical analysis.[1] In discussing the

* Economist, First National City Bank. In preparing this paper and the dissertation upon which it is based, I have benefitted from the comments and assistance of a large number of people. I am particularly indebted to Milton Friedman for key suggestions in the dissertation stage of my work and to Arthur Gandolfi for valuable discussions of various drafts of this paper. I also want to thank W. H. Bruce Brittain, Phillip Cagan, Robert J. Gordon, Wallace Huffman, Dudley D. Johnson, James P. Smith, Richard Zecher, and the managing editor of this *Review*.

[1] Milton Friedman and Anna Schwartz (1970) provide a strong statement of the position that choosing a definition of money is primarily an empirical issue. They argue further that stability in demand is an appropriate criterion for deciding between alternative definitions of money. A related criterion for defining money has been advocated by V. Karuppan Chetty, who has attempted to estimate directly the substitution relationships between various monetary totals.

factors affecting the stability of high-powered money relative to deposit-inclusive definitions of money and in the subsequent statistical analysis, I assume a simple and widely used money demand function that is derivable from the modern quantity theory and has the form:

$$(1) \qquad m = f(y_p, r)$$

where $m$ is the desired ratio of money to income, $y_p$ is permanent real income, and $r$ is "the" nominal rate of interest.

As a first approximation for statistical purposes, I also use a more restrictive formulation based upon the constant velocity model of the earlier quantity theory,

$$(2) \qquad m = k$$

where $k$—the Cambridge $k$—is a constant.

Before we actually use these demand functions to evaluate competing definitions of money empirically, I want to discuss in detail some of the other omitted influences on money demand and their likely effects on the various monetary totals. But first, let us briefly outline a more general approach to the demand for money that relaxes some of the restrictive assumptions implicit in equation (1).

A model of money demand like Benjamine Klein's (1970, 1974), which is similar to the formulations used by Edgar Feige (1964) and by Morris Perlman and is in the spirit of the new approaches to consumer theory developed by Gary Becker and Kelvin Lancaster (1966, 1971), is a particularly useful starting point. In his model Klein makes explicit the distinction between the services individuals receive from money and the real stock of money that they hold, and he assumes that the services are what provide utility to money holders. Then by specifying an individual's demand and production functions for monetary services he derives an individual's (stock) demand function for money that we can express as

$$(3) \qquad m = g(y_p, P_M, \beta_M)$$

where $P_M$ is the price of the monetary services from the marginal unit of money, equal to the difference between $r$ and the rate of interest paid on money $r_M$, and $\beta$ is an index of the quality of money, a term from the individual's production function for monetary services.[2]

### B. *Conventional Definitions of Money*

The estimation of an equation based on (1) rather than on the more complete formulation (3) can lead to several types of specification error when money is defined conventionally to include deposits.[3] Let us consider the effect of omitting $r_M$ first.

Since $r_M$ is likely to move together with the overall level of interest rates, the coefficient in the auxiliary regression of $r_M$ on $r$ will be positive, which means that the coefficient of $r$ in the regression based on (1) will be biased downwards in absolute value. But if the correlation between $r$ and $r_M$ is high, this type of specification error may have only a small effect on the stability of the demand for conventionally defined money.

Both across countries and within some countries over time this is unlikely to be the case. With no interest paid on currency, $r_M$ will be equal to the product of $r_D$, the rate paid on deposits, and $D/M$, the ratio of deposits to money. If the interest rate paid on deposits is competitive (in the sense of reflecting a zero marginal profit), it in turn will equal $r_I(1-R/D)$,

[2] Equation (3) is derivable from a demand function for monetary services of the form $S^d = D(R, y_p)$ and from a production function for monetary services of the form $S = \beta_M(M/P)$, where $S$ represents monetary services; $M/P$ is the real stock of money; $\beta_M$ is a variable coefficient of production; $R$ is the rental price of a unit of monetary services, equal to $P_M/\beta_M$; and the superscript $d$ signifies quantity demanded.

[3] See Henri Theil, pp. 504–56, for a discussion of specification analysis. Note that to simplify the discussion throughout, I am only considering one alternative definition of money, the sum of currency and total commercial bank deposits.

where $r_I$ is the interest earned on bank assets and $R/D$ is the ratio of bank reserves to deposits.[4] Hence, even if movements in $r$ and $r_I$ are closely related, differences in $R/D$ due to differences in reserve requirements or in other factors affecting banks' preferences for reserves relative to deposits will be reflected in differences in $r_D$. The existence of ceilings on either $r_I$ or directly on $r_D$, which cannot be evaded costlessly, will also reduce the correlation of $r$ and $r_M$.

What may be an even more serious problem empirically than unaccounted differences in own interest rates on money are differences in the quality of money. These differences can stem from differences in the degree of financial development of different countries or from the method used by banks to circumvent regulatory constraints on the interest they pay on deposits and earn on their assets. This second source of problems would be relevant both across countries that have differing regulations and within countries having high and variable rates of inflation along with regulations.

We can categorize these quality differences under a number of headings, each of which has slightly different implications. One assumption is that quality differences, or differences in the services offered in connection with money, are exactly equivalent to implicit interest payments on deposits. In this case their omission produces the same results as omission of explicit interest payments.

Another assumption is that quality differences are of a factor-augmenting form, involving differences in the $\beta_m$ of equation (3). In this case their impact is more difficult to assess, since an increase in $\beta_M$ has both a negative production and a positive price effect on the derived demand for money, with the net outcome ambiguous. If the net effect is positive this type of quality difference will cause a downward bias in the income coefficient in regression counterparts of (1), since countries that are more financially sophisticated are also likely to be on the whole more developed. It may also cause a downward bias (in absolute value) in the interest rate coefficient if deposit rate ceilings are the rule, since in countries with high rates of inflation banks will make more attempts to circumvent regulations.

When circumvention of interest ceilings provides the impetus, another form of what we have termed "quality changes" may take place. Banks may introduce a new liability that is unregulated as a substitute for the regulated deposits. In some instances, such as with Eurodollar deposits at foreign branches of American banks, these new liabilities may be close substitutes with existing deposits and, in principle, simply could be added to these totals to preserve their homogeneity. In actuality, however, this is unlikely to be an option, since the success of introducing new unregulated liabilities as substitutes for existing deposits depends upon their going unnoticed by the authorities. Hence, to the extent that this is the method of circumvention, the interest coefficient will be biased upwards in absolute value. Countries with high rates of inflation will have recorded deposit totals that are lower than the true totals that include the new goods.

A fourth and perhaps most realistic assumption is that quality differences affect not only the monetary service stream received from deposits, but also their substitutability with other assets. That is, we can view deposits as supplying more than one type of service, bond-type services in addition to monetary services. Some types of regulations and some patterns of finan-

[4] This formulation is discussed in Klein (1970, 1974). As he points out, even where entry into banking is regulated, competitive interest rates could still exist if competition is allowed in the provision of nonprice services which are viewed by individuals as equivalent to explicit interest payments.

cial development may lead to a change in the mix of these services. For instance, to circumvent ceilings on deposit rates, banks may offer relatively more bond-type services in connection with deposits, thus increasing their substitutability with bonds and decreasing their substitutability with currency. The end result would be to decrease the stability of the demand for money defined to include deposits.

### C. *High-Powered Money*

All of these factors—changes in own interest rates and the various types of quality changes—have direct effects on deposits but not on currency held by the public, making currency a potentially attractive alternative to broader definitions of money.[5] However, if we use currency alone as the definition of money rather than, say, the sum of currency and total commercial bank deposits or $M_2$, a consistent accounting practice would be to redefine the currency component of $M_2$ to include all of high-powered money.[6]

The use of high-powered money as the definition of money can also be viewed from a different perspective. When no interest is paid on either the deposits of commercial banks with the monetary authorities or on currency and hence on vault cash, reserve holdings of commercial banks set an upper limit on the extent to which interest can be paid on deposits. Reserves are in effect the portion of deposits which always yield only monetary services and no interest. The remainder of deposits, equal to the difference between deposits and reserves, is the portion of deposits which, at least potentially, yields interest. A holder of deposits is thus a participant in a tie-in sale, receiving these two components of deposits. The similarities in service flows between the reserve portion of deposits and currency held by the nonbank public and between the excess of deposits over reserves and other interest-bearing assets, such as bonds, may mean that the reserve portion of deposits is a close substitute with currency held by the nonbank public, while the remainder of deposits is a close substitute for the other interest-bearing assets.[7]

The analogy between this approach and Lancaster's is readily perceivable. In the simple case with which we are dealing, we can distinguish three assets and two characteristics: the former are currency, deposits, and bonds; the latter are moneyness (common to currency and deposits) and bondness (common to bonds and deposits). In this formulation, Lancaster's model explains several phenomena relevant to our discussion. In particular, it provides a rationalization for the proposition that changes in the mix of characteristics of deposits towards relatively more bondness make deposits and bonds closer substitutes, since in Lancaster's model "... closeness of substitution [is] an intrinsic effect, depending on objective characteristics of goods ..." (1967, p. 67). Hence it explains why such changes will increase the stability of high-powered money relative to money defined to include deposits. As we approach the polar case in which the only part of deposits that yield monetary services is the high-powered money por-

[5] See Friedman and Schwartz (1970, pp. 142–46), for a similar argument.

[6] When deposits of commercial banks are included in the definition of money, vault cash of commercial banks and other high-powered money reserves of commercial banks are excluded to avoid double-counting. However, if commercial bank deposits were excluded a consistent procedure would be to redefine currency held by the public to include the total monetary liabilities of the monetary authorities, or high-powered money. One way to view this is that the public is now expanded to include commercial banks just as the public can be considered to include nonbank financial intermediaries when money is defined as the total of currency and commercial bank deposits.

[7] While this may not hold as a general proposition it will be more nearly the case when variations in deposit quality affect the meaning of deposits to holders, as I explain below using Lancaster's framework.

tion, progressively smaller increases in the cost of holding deposits will lead to progressively larger decreases in deposit holdings. And in the polar case itself, deposits will be a linear combination of the characteristics of the other two goods and any increase in the cost of holding deposits will cause them to be technologically inefficient (in consumption) and hence cease to exist.[8]

## II. Empirical Results

### A. *The Data and Empirical Model*

To compare the stability of the demand for high-powered money with that of other monetary totals across countries, I have used analysis of variance techniques and their multiple regression analogues. The data used span 40 countries over the period 1952 to 1966. For each country, I have fifteen yearly observations for real per capita net national product (in 1958 dollars), for the rate of increase of the cost of living index, and for four monetary totals: high-powered money, currency held by the public, deposits, and the sum of currency and deposits, which I refer to as $M_4$ throughout. Deposits are defined broadly to include all private deposits at bank-type financial intermediaries, and high-powered money is defined as the total monetary liabilities of the monetary authorities. It therefore includes total currency outstanding along with bank and, where they exist, other private deposits at the central bank. For 22 of the countries I also have yearly figures for long-term bond yields.[9] Separate analysis is carried out for this 22-country sub-sample.

The regressions are based on the two simple general demand functions presented above and take the following form in the covariance analyses:

$$\text{(4)} \quad \log(M/Y)_{it} = \alpha_i + \delta_t + \beta_1 \log(y/L)_{it} + \beta_2 r_{it} + \epsilon_{it}$$

where $i$ denotes the country, t denotes the time period, $M/Y$ is the number of weeks of income held as the particular asset, $y$ is real income, $L$ is population, $r$ is the rate of interest, $\alpha_i$, $\delta_t$, $\beta_1$, and $\beta_2$ are parameters to be estimated, and $\epsilon_{it}$ is the error term assumed spherical normal.[10] This formulation assumes that: (a) actual and desired money balances are equal; (b) the demand function is homogeneous of degree one in population and of degree zero in prices; (c) the income elasticity is constant; (d) the semilogarithmic interest rate slope term is constant; and (e) measured income is a close approximation for permanent income. The first four assumptions are standard in much money demand analysis. The last assumption is an empirical expedient that may be imperfect within countries over time but appears reasonable

[8] In this case (which is the case of the tie-in sale) the demand for high-powered money will be unaffected by individual's choices of deposits versus currency and will depend only on their choice of characteristics. This becomes less true as deposits become more efficient, that is, as the production possibilities curve becomes more concave.

[9] The income total for almost all countries was net national product in current prices, taken from the United Nations' (*U.N.*) *Yearbook of National Account Statistics, 1969*. Publications of the *U.N.* principally issues of the *Demographic Yearbook*, provided midyear estimates of population. The bulk of the data for the four monetary totals, the data on interest rates, and the cost of living indices (1958 = 100), used to convert the income figures to constant prices and to compute yearly rates of change of prices, came from the 1966–67 and 1967–68 supplements to the International Monetary Fund's (*IMF*) *International Financial Statistics*, together with various monthly issues of that publication. Exchange rates for 1958, adjusted for changes in purchasing power parity and used to convert real per capita income to dollars, came from the *U.N. Yearbook of National Account Statistics, 1963*, Table 3B. I initially used only a broad definition of deposits because of difficulties in constructing a series for demand deposits at commercial banks, the most important of which was that for some years for some countries separate demand deposit data were not available from the *IMF*. As reported in the conclusion, later work with a demand deposit series obtained in part from other sources alters none of my results.

[10] The corresponding regression for an analysis of variance imposes the constraint that $\beta_1 = \beta_2 = 0$.

TABLE 1—ANALYSES OF VARIANCE AND COVARIANCE[a]

| Mean Squares for: | Degrees of Freedom | Variable[b] *log* $(H/Y)$ | *log* $(C/Y)$ | *log* $(D/Y)$ | *log* $(M_4/Y)$ |
|---|---|---|---|---|---|
| **40-Country Sample** | | | | | |
| Analysis of Variance | | | | | |
| Countries | 39 | 1.364 | 2.156 | 8.512 | 5.430 |
| Years | 14 | .031 | .136 | .546 | .179 |
| Remainder | 546 | .015 | .009 | .029 | .016 |
| Analysis of Covariance | | | | | |
| Countries | 39 | 1.239 | 2.086 | 3.011 | 2.235 |
| Years | 14 | .018 | .016 | .069 | .016 |
| Remainder | 545 | .015 | .009 | .028 | .015 |
| **22-Country Sample** | | | | | |
| Analysis of Variance | | | | | |
| Countries | 21 | .801 | 1.892 | 6.369 | 3.777 |
| Years | 14 | .059 | .073 | .099 | .028 |
| Remainder | 294 | .011 | .009 | .015 | .008 |
| Analysis of Covariance | | | | | |
| Countries | 21 | .760 | 1.828 | 1.178 | .878 |
| Years | 14 | .012 | .002 | .009 | .004 |
| Remainder | 293 | .011 | .008 | .012 | .007 |

[a] The International Monetary Fund (*IMF*) was the main source of monetary data and the United Nations (*U.N.*) the main source of income data.

[b] High-powered money, currency, broadly defined deposits, and the sum of currency and broadly defined deposits are represented by $H$, $C$, $D$, and $M_4$, respectively, and nominal income by $Y$.

across countries, where differences in transitory income seem to be a relatively small proportion of differences in measured income. The next to last assumption, unlike the logarithmic formulation that is also sometimes used, implies a finite demand for money at a zero rate of interest, and is justifiable if there are some increasing marginal costs of holding money.

### B. *Stability of Alternative Totals*

Analyses of variance and of covariance of the monetary totals are presented in Table 1 to provide evidence on the relative stability of each of the four totals over time and across countries under both specifications of the demand for money function. The mean squares for countries and years of the analyses of variance are measures of the absolute variation in the fraction of income held in the form of each asset ( or in velocity) across countries and over time, while those in the analyses of covariance are measures of the variations in asset holdings across countries and over time after allowance is made for the effects of the independent variables, real per capita income in the full sample, and both real per capita income and the rate of interest in the subsample. Tests of significance of these mean squares are therefore tests for significant departures from the respective models of the demand for money.[11]

[11] The mean square for countries is the regression sum of squares due to the dummy variables used to allow the $\alpha_i$ to differ, the mean square for years is the regression sum of squares due to the dummy variables used to allow the $\delta_t$ to differ, and the remainder sum of squares is the sum of squared errors from the regression. A test of the null hypothesis that the $\alpha_i$ are all equal is a test

The results of both types of analyses strongly support the view that narrower totals in general and high-powered money in particular are more homogeneous across countries. In the analyses of variance for the full sample, the country mean square for high-powered money is less than two-thirds that of currency and less than one-third of the other two assets. In the subsample, it is at the most one-half of the mean square of the other three assets. In the analyses of covariance, high-powered money, too, is the least variable across countries in both samples. What is particularly surprising when these results and the analyses of variance results are compared is the minute reductions in the mean square for high-powered money across countries (1.36 to 1.24 in the full sample and .80 to .76 in the subsample) and the lower unadjusted mean square for high-powered money than the adjusted mean squares for either deposits or $M_4$ (1.36 versus 3.01 and 2.24 in the full sample and .80 versus 1.18 and .88 in the subsample).

Over time the results are more mixed, and appear somewhat different from those obtained across countries. In the analyses of variance, high-powered money proves least variable in the full sample and $M_4$ in the small sample, while in the analyses of covariance, currency and $M_4$ prove equally less variable in the full sample and currency proves least variable in the small sample. However, in both sets of covariance analyses, with the exception of deposits in the full sample, these mean squares for years are insignificant at the .05 level.[12]

for no significant variation across countries; a test of the null hypothesis that the $\delta_j$ are all equal is a test for no significant variation over time. Rejection of the alternative hypotheses that in each instance they are unequal implies that velocity can be taken as constant and equal to its geometric mean.

[12] All of the mean squares in the analyses of variance, with the exception of the mean square for years for high-powered in the full sample (which is barely significant at .05), are highly significant. Those for countries in the analyses of covariance remain so.

## C. *Estimated Demand Functions*

Summary statistics and coefficients of the independent variables for the regressions underlying the covariance analysis—pooled regressions with individual intercepts for countries and years—together with the results of five other types of regressions are presented in Tables 2 and 3. These results permit comparison with the findings of other demand for money studies and also help illuminate the differences between the cross-country and within-country stabilities of different totals.

The implied income elasticities and the interest rate coefficients are for the most part highly significant and agree closely with estimates obtained in previous money demand studies.[13] The pooled regressions with single intercepts, the pooled regressions with individual yearly intercepts, and the country-mean regressions, all capture primarily cross-country relationships.[14] In these regressions the income elasticities of demand for real balances range from close to 1.0 for currency to close to 1.5 for deposits. The interest elasticities (evaluated at the mean of the interest rate, .054) all indicate an inelastic response of asset holdings to changes in interest rates and range from less than $-.10$ for high-powered money to a little over $-.40$ for money.

[13] For example, see the time-series studies by Klein, Allan Meltzer, Friedman and Schwartz (forthcoming), A. A. Walters, and the cross-state studies by Feige and Arthur Gandolfi. Note that my estimates of income elasticities (of real cash balances held in the form of each asset) equal one plus the income coefficient. Hence an income coefficient significantly different from minus one implies an income elasticity significantly different from zero. Correspondingly, a coefficient of determination close to zero implies no additional gain in explanatory power from relaxing the assumption of unit income elasticity (or from including interest rates).

[14] Both the yearly intercepts and the averaging inherent in the country-mean regressions dictate this. The primarily cross-country nature of the regressions with single intercepts is peculiar to these data. Such regressions treat within-country and cross-country variation the same, and the latter is a predominant portion of the total variation in holdings.

TABLE 2—SUMMARY STATISTICS FOR REGRESSIONS FOR THE 40-COUNTRY SAMPLE[a]

| Dependent Variable | Coefficients of[b] *log* $(y/L)$ | Coefficients of[b] $g_p$ | $R^2$ | Standard Error | Coefficients of[b] *log* $(y/L)$ | $R^2$ | Standard Error |
|---|---|---|---|---|---|---|---|
| **Pooled: Intercepts for Years** | | | | | **Pooled: Intercepts for Countries** | | |
| *log* $(H/Y)$ | .090 | | .08 | .310 | −.117 | .02 | .122 |
| | (.013) | | | | (.033) | | |
| *log* $(C/Y)$ | .070 | | .03 | .384 | −.374 | .28 | .094 |
| | (.016) | | | | (.026) | | |
| *log* $(D/Y)$ | .497 | | .62 | .477 | .745 | .32 | .170 |
| | (.019) | | | | (.046) | | |
| *log* $(M_4/Y)$ | .455 | | .57 | .404 | .455 | .25 | .122 |
| | (.016) | | | | (.033) | | |
| **Pooled: Single Intercept** | | | | | **Pooled: Intercepts for Countries and Years** | | |
| *log* $(H/Y)$ | .085 | | .07 | .309 | .009 | .00 | .122 |
| | (.012) | | | | (.053) | | |
| *log* $(C/Y)$ | .062 | | .02 | .385 | −.217 | .05 | .093 |
| | (.016) | | | | (.041) | | |
| *log* $(D/Y)$ | .601 | | .62 | .473 | .417 | .05 | .167 |
| | (.019) | | | | (.074) | | |
| *log* $(M_4/Y)$ | .456 | | .57 | .400 | .332 | .06 | .122 |
| | (.016) | | | | (.050) | | |
| **Country Means** | | | | | **Pooled: Error Component for Countries** | | |
| *log* $(H/Y)$ | .086 | −.214 | .05 | .295 | .090 | .09 | .274 |
| | (.048) | (.574) | | | (.012) | | |
| *log* $(C/Y)$ | .049 | −1.164 | .05 | .370 | .072 | .04 | .355 |
| | (.060) | (.720) | | | (.015) | | |
| *log* $(D/Y)$ | .586 | −.587 | .63 | .457 | .599 | .65 | .427 |
| | (.063) | (.892) | | | (.018) | | |
| *log* $(M_4/Y)$ | .438 | −.871 | .58 | .390 | .456 | .59 | .368 |
| | (.062) | (.760) | | | (.016) | | |

[a] The *IMF* was the main source of monetary and price data and the *U.N.* of income and exchange rate data.
[b] Real per capita income in 1958 dollars is represented by $y/L$ and the average rate of change of the cost of living index by $g_p$. Standard errors of the coefficients are beneath them in parentheses.

The pooled regressions with individual country constants and the pooled regressions with individual yearly and country constants both seem to capture primarily time-series relations. They yield estimated elasticities fairly similar to those from the three types of cross-country regressions. The only substantial differences are in the interest elasticities of deposits and money and the income elasticities of high-powered money and currency, all of which are lower in the pooled regressions with individual country constants than in the cross-country regressions.

The standard errors of these regressions provide information similar to though not fully independent of the results of the covariance analyses. Again, high-powered money is the most stable total across countries. Within countries currency appears somewhat more stable than alternative totals, though again the differences among them are slight.

A problem with the within-country comparisons, however, is that for none of the assets are the individual within-country regressions statistically homogeneous.[15] The implication is that from the standpoint of the time-series relations there are

[15] The $F$-ratio to test jointly the homogeneity of both slopes and intercepts in the regressions for the individual countries, showed significant differences at less than the .001 level for all assets in both samples. Tests of individual yearly regressions revealed the opposite—no significant differences at the .10 level.

TABLE 3—SUMMARY STATISTICS FOR REGRESSIONS FOR THE 22-COUNTRY SAMPLE[a]

| Dependent Variable | Coefficients of[b] $log\ (y/L)$ | $r$ | $R^2$ | Standard Error | Coefficients of[b] $log\ (y/L)$ | $r$ | $R^2$ | Standard Error |
|---|---|---|---|---|---|---|---|---|
| **Pooled: Intercepts for Years** | | | | | **Pooled: Intercepts for Countries** | | | |
| $log\ (H/Y)$ | .044 | −1.143 | .04 | .248 | −.206 | −3.254 | .17 | .107 |
| | (.013) | (.734) | | | (.044) | (.805) | | |
| $log\ (C/Y)$ | −.016 | −3.795 | .04 | .361 | −.336 | −3.468 | .36 | .089 |
| | (.019) | (1.068) | | | (.037) | (.671) | | |
| $log\ (D/Y)$ | .523 | −8.065 | .80 | .301 | .577 | −2.199 | .35 | .110 |
| | (.016) | (.889) | | | (.046) | (.828) | | |
| $log\ (M_4/Y)$ | .379 | −7.891 | .75 | .255 | .354 | −2.731 | .26 | .081 |
| | (.013) | (.756) | | | (.034) | (.614) | | |
| **Pooled: Single Intercept** | | | | | **Pooled: Intercepts for Countries and Years** | | | |
| $log\ (H/Y)$ | .039 | −1.745 | .05 | .248 | −.007 | −1.769 | .01 | .106 |
| | (.013) | (.710) | | | (.073) | (.923) | | |
| $log\ (C/Y)$ | −.020 | −4.214 | .05 | .355 | −.331 | −3.488 | .13 | .090 |
| | (.018) | (1.018) | | | (.062) | (.784) | | |
| $log\ (D/Y)$ | .527 | −7.568 | .79 | .297 | .495 | −2.993 | .17 | .110 |
| | (.015) | (.851) | | | (.075) | (.957) | | |
| $log\ (M_4/Y)$ | .381 | −7.568 | .75 | .251 | .296 | −3.280 | .16 | .082 |
| | (.013) | (.721) | | | (.056) | (.712) | | |
| **Country Means** | | | | | **Pooled: Error Component for Countries** | | | |
| $log\ (H/Y)$ | .045 | −1.069 | .00 | .236 | .045 | −1.038 | .05 | .213 |
| | (.048) | (2.900) | | | (.012) | (.696) | | |
| $log\ (C/Y)$ | −.014 | −3.897 | .00 | .366 | −.014 | −4.015 | .04 | .328 |
| | (.074) | (4.400) | | | (.018) | (1.072) | | |
| $log\ (D/Y)$ | .524 | −8.747 | .82 | .291 | .524 | −8.746 | .82 | .263 |
| | (.059) | (3.577) | | | (.014) | (.857) | | |
| $log\ (M_4/Y)$ | .380 | −8.523 | .77 | .251 | .380 | −8.543 | .23 | .226 |
| | (.051) | (4.318) | | | (.012) | (.739) | | |

[a] The *IMF* was the main source of monetary, price, and interest rate data and the *U.N.* of income and exchange rate data.

[b] Real per capita income in 1958 dollars is represented by $y/L$ and the rate of interest by $r$. Standard errors of the coefficients are beneath them in parentheses.

some important omitted variables that differ in their impact from one country to the next. An alternative way of dealing with the time-series relations therefore may be in order. One solution is to use an error component model, such as that proposed by Pietro Balestra and Marc Nerlove, which views the error term in the model as being made up of two parts: a cross-country error and a remainder. When this technique is used we get another set of time-series estimates which are also presented in Tables 2 and 3. What they show is, as in the cross-section regressions, the clear superiority of high-powered money. In both samples the standard errors of estimate for the regressions for high-powered are lower than those for the regressions for the other assets.

One reason for the differences between the relative stabilities of different totals across countries and within countries uncovered in some of the regressions is brought out by examination of the interest rate coefficients. In the regressions in which high-powered money is most stable, its sensitivity to interest rates is much lower than that of deposits and $M_4$. This is precisely what one would expect if there were specification bias in the deposit and

money regressions, caused by either the existence of close substitutes for deposits introduced to circumvent interest rate ceilings, or by a greater degree of substitutability of deposits for bonds than of the high-powered money portion of deposits for bonds.

What may be more important sources of differences between the two sets of results are differences between cross-country and within-country variations in deposit quality. One might plausibly expect the former to be considerably greater since differences in the degree of financial development across countries are probably much greater than those that are likely to exist within countries over the relatively short time span covered by the sample. Omitting this variable, therefore, would decrease the stability of $M_4$ relative to narrower totals across countries and have little effect within countries.

This second explanation, moreover, is given credence by several additional types of evidence. In the analyses of variance and of covariance, variations in holdings of deposits and of $M_4$, both absolutely and relative to high-powered money, were much lower in the small sample that is the more homogeneous of the two with respect to degree of financial development. Regressions run with country means and incorporating the ratio of banking offices to population as an index of deposit quality also support this hypothesis.[16] Hence, differing degrees of financial sophistication seem to account for at least part of the difference between the performance of $M_4$ and high-powered money across countries and, therefore, may also account for part of the differences between the cross-country and the time-series results.

## III. Summary and Conclusions

The central theoretical proposition of this paper is that substantial variations in the cost of holding deposits relative to the costs of holding currency and other assets and in the quality or "moneyness" of deposits, such as exist across countries and within some countries having high and varying rates of inflation and ceilings on deposit rates, will substantially reduce the stability of the demand for money defined conventionally to include deposits while having only a minor effect on the stability of the demand for money defined more narrowly as high-powered money alone. In these circumstances high-powered money may therefore be a more useful definition of money than deposit-inclusive monetary totals.

The empirical evidence I have presented in the main supports this proposition. I have found that across countries high-powered money is the most stable of the four totals I have examined even when it is judged on the basis of the simple constant

[16] The data and their sources are described in my dissertation. The regression for the country means of deposits for 34 countries in the full sample for which data were available was

$$\log (D/Y) = .413 + \underset{(.093)}{.345} \log (y/L) + \underset{(563.3)}{1637.8} B/L - \underset{(.776)}{1.260} g_p$$

$R^2 = .74$; $SE = .382$

and for the 22 countries in the small sample was

$$\log (D/Y) = .484 + \underset{(.076)}{.425} \log (y/L) + \underset{(492.7)}{945.0} B/L - \underset{(3.347)}{9.174} r$$

$R^2 = .85$; $SE = .274$

where $B/L$ is the deposit quality variable and where standard errors are beneath the coefficients in parentheses. Higher income coefficients were obtained in similar regressions that omitted $B/L$ (.528 and .524, respectively), which is what one would expect if an increase in deposit quality caused by increased financial development has a positive effect on the demand for deposits and if more financially developed economies are also more developed on the whole. In the above regressions $B/L$ was significant in both instances at the .05 level which it never was when I used high-powered money as the dependent variable.

velocity model and the other totals are judged on the basis of a more sophisticated model of money demand. Moreover, this greater stability of high-powered money appears directly attributable to the factors which theory suggests will have an important destabilizing influence on the demand for deposits. Specification analysis of various types of regressions, the observed differences in the stability of $M_4$ relative to high-powered money in the two samples, and the results of regressions which use a proxy for the quality of deposits as an independent variable all tend to confirm the importance of variations in deposit quality. The one major piece of evidence which does not support the hypothesis is that high-powered money is if anything less stable than either currency or $M_4$ within countries. But as already mentioned, this finding is consistent with the other evidence if, as appears reasonable, variations in deposit quality are on average of only minor importance within countries.

Moreover, this greater cross-country stability of narrower totals in general and high-powered money in particular is not just a statistical quirk peculiar to these data. Analysis of other studies indicates that this result still holds when other assets, specifically $M_1$ and its deposit component, are considered and other samples are used.[17] As an additional check, I compiled an $M_1$ series for my group of countries. The regressions I ran for both the 40 countries and the 22 countries also indicate that high-powered money is more stable than either $M_1$ or the deposit component of $M_1$.[18]

My findings also provide evidence of several sorts bearing on the overall stability of money demand. They show that given the variations in factors which could plausibly be expected to affect the demand for money across countries, it is reasonably stable. Estimates of international variations in money holdings are not radically greater than the intranational estimates.[19] In addition, even though these international variations are both significant and substantial, it is clear that even so simple an hypothesis as constant velocity has considerable merit. For example, from knowledge of high-powered money and of its average velocity, one can account for an overwhelming proportion of the variance in the level of nominal income or in its rate of change— over 95 percent of the variance of the level in 1958 and over 90 percent of the variance of its average annual rate of change. In addition, my estimates of money demand functions on the whole accord fairly well with esti-

[17] Estimates of coefficients of variation of velocity derived from regressions similar to mine contained in studies by Hannan Ezekial and Joseph Adekunle, by Morris Perlman, and by Henry Wallich show that currency is more stable than broader totals (either $M_4$ or both $M_1$ and $M_4$, depending upon the study) and that holdings of both broader totals are more variable than high-powered money is in my full sample. Additional comparisons which I have made with an extended and otherwise revised version of Perlman's data and which I report in my dissertation show high-powered money to be the most stable total.

[18] The regression for the country means of $M_1$ holdings for the 40-country sample was

$$log\ (M_1/Y) = .840 + \underset{(.056)}{.259}\ log\ (y/L) - \underset{(.411)}{.276}g_p$$

$R^2 = .39; \quad SE = .345$

and for the 22-country sample was

$$log\ (M_1/Y) = 1.337 + \underset{(.056)}{.233}\ log\ (y/L) - \underset{(3.338)}{3.912}r$$

$R^2 = .50; \quad SE = .278$

where standard errors of the coefficients are beneath them in parentheses. We can compare the standard errors of estimate of these regressions of .345 and .278 with the standard errors of estimate for similar regressions for high-powered money of .295 and .236, respectively.

[19] Friedman and Schwartz (forthcoming) report coefficients of variation of the levels of velocity of $M_2$ in the United States of .32, and in the United Kingdom of .18 over the years 1880–1968. Estimates for the velocity of total adjusted commercial bank deposits from cross-state data, made available by Gandolfi, range from .17 in 1929 to .20 in 1933.

mates others have obtained from long-term studies of the United States and the United Kingdom and from cross-section studies of the United States.[20] Given the substantial independence of my data from theirs, this close agreement of results provides a strong corroboration of their results. Conversely, it casts considerable doubt on estimates from postwar time-series for those two countries which in general conflict greatly with both long-term time-series and cross-section estimates.[21]

An obvious implication of my findings is that for international studies in which the stability of the demand for money is a key assumption, such as in monetarist models of the balance of payments or of exchange rates, the use of high-powered money as the definition of money may prove extremely fruitful. In studies of that type it also has the added advantage of eliminating the need for separate relationships to explain the conventional money multiplier and conventional money demand. For countries having no income data or data of dubious accuracy there are also direct applications for these findings. By extension they also may prove useful in time-series studies of economic conditions in countries experiencing rapid inflation or substantial changes in their financial structure.

My results also have implications for monetary policy. Since inflation coupled with regulation of banking tends to reduce the homogeneity of deposits, conventional definitions of money may become highly imperfect indicators for monetary policy in such situations. On the one hand, my findings suggest that policy should be such that supply and demand are not made more interdependent. And on the other, they suggest that, having made them so, the monetary authorities should look for a more homogeneous total as an indicator and not deduce from the instability in demand of a nonhomogeneous total that monetary aggregates play only a weak role in the economy.[22]

Perhaps the most interesting implications of these results are for the definition of money. Most of the current debate over how to define money has centered upon whether to include various types of time and savings deposits in the definition of money. My results suggest that the focus of current debate has been too narrow, that the range of plausible alternative definitions of money and the factors influencing the selection of one of these alternatives are both broader than is commonly realized. They indicate that an earlier definition of money as high-powered money alone may be a useful alternative to these conventional definitions of money. And because the differences in the performance of high-powered money and deposit-inclusive definitions of money across countries can be attributed to variations in deposit quality, these results also indicate the importance of asset quality in deciding upon a definition of money.

[20] In addition to the studies cited in fnn. 13 and 17, see the cross-country study by James Hanson and Robert Vogel.

[21] Estimates of income elasticities obtained from the studies referred to in the previous footnote tend to fall in the range of 1.0 to 1.3. Estimates obtained with postwar data (see Stephen Goldfeld) are usually considerably below 1.0.

[22] The United Kingdom currently provides an instructive example. Changes in the regulations surrounding banking appear to be a major cause of the marked divergences over the past several years in the growth rates of the two published deposit-inclusive definitions of money. These changes have undoubtedly decreased the homogeneity of both totals and thus have drastically increased the difficulties in assessing British policy.

## REFERENCES

**P. Balestra and M. Nerlove**, "Pooling Cross Section and Time-Series Data in the Estimation of a Dynamic Model: The Demand for Natural Gas," *Econometrica*, July 1966, *34*, 585–612.

G. S. Becker, *Economic Theory*, New York 1971.

V. K. Chetty, "On Measuring the Nearness of Near-Moneys," *Amer. Econ. Rev.*, June 1969, *59*, 270–81.

H. Ezekial and J. O. Adekunle, "The Secular Behavior of Income Velocity," *Int. Monet. Fund Staff Pap.*, July 1969, *16*, 224–37.

E. Feige, *The Demand for Liquid Assets: A Temporal Cross-Section Analysis*, Englewood Cliffs 1964.

———, "Alternative Temporal Cross-Section Specifications of the Demand for Demand Deposits," in H. G. Johnson and A. R. Nobay, eds., *Issues in Monetary Economics*, London 1974.

M. Friedman and A. J. Schwartz, *Monetary Statistics of the United States, Estimates, Sources and Methods*, New York 1970.

——— and ———, *Monetary Trends in the United States and the United Kingdom*, Nat. Bur. Econ. Res. forthcoming.

A. E. Gandolfi, "The Stability of the Demand for Money During the Great Contraction, 1929–1933," *J. Polit. Econ.*, Oct. 1974, *84*, 969–83.

S. M. Goldfeld, "The Demand for Money Revisited," *Brookings Papers*, Washington 1973, *3*, 577–646.

J. S. Hanson and Robert C. Vogel, "Inflation and Monetary Velocity in Latin America," *Rev. Econ. Statist.*, Aug. 1973, *55*, 365–70.

H. S. Houthakker, "New Evidence on Demand Elasticities," *Econometrica*, Apr. 1965, *33*, 272–88.

B. Klein, "The Payment of Interest on Commercial Bank Deposits and the Price of Money," unpublished doctoral dissertation, Univ. Chicago 1970.

———, "Competitive Interest Payments on Bank Deposits and the Long-Run Demand for Money," *Amer. Econ. Rev.*, Dec. 1974, *64*, 931–49.

K. Lancaster, "A New Approach to Consumer Theory," *J. Polit. Econ.*, Apr. 1966, *77*, 132–57.

———, *Consumer Demand: A New Approach*, New York 1971.

J. R. Lothian, "The Demand for High-Powered Money," unpublished doctoral dissertation, Univ. Chicago 1973.

A. H. Meltzer, "The Demand for Money: The Evidence from the Time-Series," *J. Polit. Econ.*, June 1963, *71*, 129–246.

M. Perlman, "International Differences in Liquid Asset Portfolios," in D. Meiselman, ed., *Varieties of Monetary Experience*, Chicago 1970.

H. Wallich, "Quantity Theory and Quantity Policy," *Ten Economic Studies in The Tradition of Irving Fisher*, New York 1967.

A. A. Walters, *Money in Boom and Slump*, 3d ed., Inst. Econ. Anal., Hobart Paper, *44*, London 1971.

International Monetary Fund, *International Financial Statistics*, 1952–1969.

United Nations, *Demographic Yearbook*, 1965–69.

———, *Statistical Yearbook*, 1961–67.

———, *Yearbook of National Account Statistics*, 1963, 1969.

# Benefit Shares and Majority Voting

*By* ARTHUR T. DENZAU AND ROBERT J. MACKAY*

In the burgeoning literature on collective decision making, attention generally has been focused on the public provision of goods in equal quantities to all recipients. As a result, there has been little positive analysis of the effects of differential distribution of the benefits of public activities. This state of affairs contrasts with the more developed literature on the effects of differential distribution of the costs of public activities. By analogy to the tax share concept in the study of cost sharing, this paper examines the effects of differential benefit shares, or of changing benefit shares, on voting behavior at the individual and aggregate level in a majority voting model.

Just as the early literature on cost sharing often assumed equal sharing by all and thus the same tax share for all, the assumption of equal quantities for all can be interpreted as equal benefit shares. This equal quantities assumption requires a distinction between production units (the units relevant for production and cost functions) and consumption units (the units relevant for individual preference relations).[1] This distinction immediately raises the issue of the distribution of public benefits, i.e., the transformation of production units into consumption units. A benefit share measures the rate at which a production unit is transformed into an individual's consumption units.[2] For example, in the case of a "pure public good," without exclusion, the benefit shares would all equal unity.

The problem this paper is concerned with can now be stated in its simplest form by the following example. Suppose that the state has made private trading in bread illegal. In lieu of market exchange, the government distributes the bread at a zero price, with the cost of the bread met through taxation. The total amount of the bread provided (units in production) is set by a majority vote, while the distribution of bread (units in consumption) is based on a fixed share arrangement. For example, if the total distribution is 100 loaves per week and Joe's benefit share is .02, then he receives 2 loaves per week. If the distribution scheme is changed—the fixed benefit shares are changed—how does this affect the total amount of bread (in production units) that voters will most prefer? On what factors does each voter's demand depend and how do these individual demands aggregate into market level results?

Section IA examines the individual choice problem: how does an individual's most preferred public sector size change as his benefit share changes? Buchanan (1968, p. 54) conjectures that an individual with

* Assistant professors, department of economics, and research associates, Center for Study of Public Choice, Virginia Polytechnic Institute and State University. We are indebted to the Ford Foundation for financial support of this research. We acknowledge the helpful comments of James Buchanan, Robert Parks and Theodore Bergstrom on earlier drafts of the paper. We are also deeply indebted to our late colleague Winston Bush who helped us launch our joint research. Earlier drafts of this paper were presented at the Southern Economic Association Meetings in Atlanta 1974, and in seminars at the University of Arizona and the University of Illinois.

[1] This distinction is made in Buchanan (1966), further developed in Buchanan (1968, pp. 52–56), and is a major issue in the discussion between Albert Breton and Buchanan (1967).

[2] Recent empirical work has explicitly utilized this distinction by assuming equal sharing in the consumption units. For examples of the use of the equal sharing assumption in the recent empirical literature, see Thomas Borcherding and Robert Deacon, Theodore Bergstrom and Robert Goodman, or Robert Spann.

a higher benefit share would prefer a larger public sector budget. The present paper establishes empirically plausible conditions under which this conjecture is false.[3] If the $i$th voter has price-inelastic (elastic) demand for the production of the publicly provided good, then an increase in his perceived benefit share $\alpha_i$ causes him to prefer and hence vote for a smaller (larger) size of the public sector and its budget. Section IB presents a simple market level example to illustrate the notion of benefit share progressivity and examine the effects of changes in the degree of progressivity on market level outcomes. Finally, Section II applies the analysis to the public provision of a purely private good, bread, and to a more typical public service, fire protection. These applications provide explicit interpretations of the meaning of benefit shares.

## I. The Basic Model

### A. *Individual Level Analysis*

Suppose that the $i$th voter's preferences can be represented by the general utility function:

$$(1) \qquad U_i = U_i(z_i, q_i)$$

where $z_i$ is the voter's consumption of the privately provided good and $q_i$ is the voter's perception of his personal consumption of the publicly provided good. In this section, the voter's perception of his personal consumption is assumed to be a linear function of the total provision of the publicly provided good $x$, i.e.,

$$(2) \qquad q_i = \alpha_i x$$

where $\alpha_i$ is the voter's perception of his benefit share.[4] Generally, the benefit share $\alpha_i$ is considered to be greater than or equal to zero, and less than or equal to unity.[5] The total provision is measured in production units while the individual only values his own consumption units $q_i$.[6]

When voting, each individual determines his most desired provision of the publicly provided good $x$ by maximizing his utility function subject to equation (2) and his budget constraint:

$$(3) \qquad y_i \geqq z_i + p_i^x x$$

where $y_i$ is his income and $p_i^x$ is the price, in terms of the numeraire good $z$, to voter $i$ of an additional unit of $x$. That is,

$$(4) \qquad p_i^x \equiv \tau_i p$$

where $\tau_i$ is the $i$th voter's nonnegative tax share and $p$ is the market price of an additional unit of $x$ in terms of the numeraire good.

By substituting equations (2) and (4) into (3), the voter's maximization problem can be stated as:

$$(5) \qquad \max_{z_i, q_i} U_i = U_i(z_i, q_i)$$

[3] In the present paper, benefit shares are not determined as they might be in a normative model such as Buchanan (1968, pp. 60–65).

[4] Equation (2) implies a system of proportional distribution of the production of $x$. This is also the meaning of Figure 4.1 of Buchanan (1968, p. 59). Much of the meaning of equation (2) can be derived from Buchanan (1968, pp. 52–65). Equation (2) is also identical in interpretation to the $T(.,.)$ map of Charles Plott and Robert Meyer (p. 21). Although they further distinguish between available amounts $q_p^i$ and appropriated units $x_k^i$, most of their discussion involves the distinction between production and consumption units.

[5] Perhaps a better terminology would be to call $\alpha_i$ a coefficient of distribution of the publicly provided production flow $x$. The obvious analogy between benefit share and tax share has a certain appeal, however, and for this reason the expression "benefit share" is used instead of, say, "coefficient of distribution."

[6] The basic function of equation (2) is to provide a definite relationship between the potentially observable public sector state $x$ and the more subjective consumption unit $q_i$, which directly enters the voter's utility function. The production of the publicly provided good $x$ is only potentially observable since the total budget size $px$ is the variable that is generally observable. In short, existing data often just contain total budget size and, hence, it may be difficult to separate quantity and price. In order to avoid the more subjective measurement problems involved in dealing with $q_i$ the attention of this paper is focused on equations dealing with $x$ rather than $q_i$.

subject to $y_i \geqq z_i + \left(\frac{\tau_i p}{\alpha_i}\right) q_i = z_i + \overset{q}{p_i} q_i$

One can treat $p_i^q$ as the price to voter $i$ of *consuming* an additional unit of the publicly provided good—it is his tax price for an additional unit of $q_i$. Under appropriate conditions this maximization problem yields a differentiable demand function of the following general form:

$$\overset{d}{q_i} = g_i(\overset{q}{p_i}, y_i) \tag{6}$$

The demand function given in (6) can be immediately converted into the $i$th voter's demand for the total provision of the publicly provided good $x_i^d$ as a function of the benefit share, tax share, the price of a unit of $x$, and income. That is,

$$\begin{aligned} \overset{d}{x_i} \equiv \overset{-1}{\alpha_i} \overset{d}{q_i} &= \overset{-1}{\alpha_i} g_i(\tau_i p / \alpha_i, y_i) \\ &= h_i(\tau_i p, \alpha_i, y_i) \end{aligned} \tag{7}$$

Voter $i$'s demand for the total provision of the publicly provided good $x_i^d$ is best interpreted as his most preferred or desired public sector size—the quantity associated with the peak of his single-peaked public sector preference relation.[7]

Equation (7) could be used to determine how the voter's demand for $x$ varies with changes in his tax share, benefit share, income, and the price of a unit of the public good ($\tau_i$, $\alpha_i$, $y_i$, and $p$, respectively). The major concern of this paper, however, is with the effects of changes in benefit shares on $x_i^d$. Differentiating equation (7) with respect to $\alpha_i$ and rearranging gives:[8,9]

$$\frac{\partial \overset{d}{x_i}}{\partial \alpha_i} = -\frac{1}{\alpha_i} \overset{d}{x_i}(1 + \eta_{x,p}) \tag{8}$$

The expression can be signed if one knows the price elasticity of demand for $x$, i.e.,

$$\frac{\partial \overset{d}{x_i}}{\partial \alpha_i} \gtrless 0 \qquad \text{as } \eta_{x,p} \lesseqgtr -1 \tag{9}$$

Equation (9) gives the fundamental result of this section for the linear perception relation given by equation (2). In words, *if voter $i$ has price-inelastic (elastic) demand for the production or total provision of the publicly provided good, then an increase in his perceived benefit share $\alpha_i$ causes him to prefer and, hence, vote for a smaller (larger) size of the public sector and budget.*

The basic mechanism for this result is quite simple and is most clearly seen by examining the demand function for $q_i$ along with equation (2). This linear share assumption implies the equality of the tax-price demand elasticities for $x$ and $q$; i.e.,

$$\eta_{x,p} = \eta_{q,p} \tag{10}$$

Now consider voter $i$'s behavior as $\alpha_i$ changes if his demand is price inelastic. In this case, a doubling of $\alpha_i$, causing a halving of $p_i^q$, leads to less than a doubling in $q_i^d$. But equation (2) implies that the doubling of $\alpha_i$ causes a doubling of the amount of the good that is actually available to the $i$th voter for consumption from the *existing* production of the publicly provided good. As a result, the $i$th voter will vote for a smaller total production of the publicly provided good $x$, attempting to bring his actual or perceived consumption of the good into equality with his desired

[7] The voter's public sector preference function provides his ordering of public sector states by assigning a utility level or index to alternate states of the public sector measured in terms of production units. The state of the public sector could also be the total budget size. The quasi concavity of the utility function and the convexity of the budget set in (5) insures the single peakedness.

[8] The elasticity $\eta_{x,p} = (\partial x_i^d / \partial p_i^x)(p_i^x / x_i^d)$, and $\eta_{q,p} = (\partial q_i^d / \partial p_i^q)(p_i^q / q_i^d)$.

[9] At this point it may help to interpret equation (8) in terms of its operational content. First, one may be able to observe the same individual having two different benefit shares: shifts in policy or jurisdictional consolidation could lead to this possibility. Secondly, one might assume that for empirical testing, all individuals have the same tastes, but different benefit shares. A final possibility would involve individuals having different tastes with the benefit shares distributed independently of the taste parameters.

consumption. By a similar argument, if the demand for the consumption of the publicly provided good is elastic, then an increase in the voter's benefit share will cause him to vote for a larger total production of the publicly provided good. Finally, if the elasticity is unitary, then the voter's desired size of the public sector is independent of his benefit share—changes in his benefit share will not affect his voting behavior.[10]

### B. *Market Level Analysis*

The analysis of individual voter behavior in the previous section must be extended to the market level if it is to provide insights into the effects of changes in benefit shares on the equilibrium size of the public sector budget. This section provides an example using the above analysis to generate market level results.

#### 1. *A Simple Model*

The present example introduces the notion of benefit share progressivity and examines the effects of changes in progressivity on market level outcomes. To develop this example, explicit functional forms are assumed for the demand functions, the tax shares, and the benefit shares.[11]

Each voter is assumed to have the following demand function for $q$:

$$q_i^d = \rho_0 y_i^{\rho_1} (p_i^q)^{\rho_2}$$

Substituting in for $q_i^d$ and $p_i^q$ yields from (2) and (6):

$$(11) \qquad x_i^d = \rho_0 y_i^{\rho_1} \tau_i^{\rho_2} p^{\rho_2} \alpha_i^{-(1+\rho_2)}$$

In this example, tax shares are determined by the following function:

$$(12) \qquad \tau_i = \frac{y_i^\beta}{\sum_j y_j^\beta} \qquad j = 1, \ldots, m$$

This formulation of the tax shares may possibly result from a system of property taxation.[12] Finally, perceived benefit shares are also assumed to depend only on income:

$$(13) \qquad \alpha_i = \frac{c y_i^\gamma}{\sum_j y_j^\gamma}$$

Note that the benefit shares sum to $c$. If $c$ equals unity, then the example may involve the case of public provision of a purely private good. The parameter $\gamma$ reflects in some sense the extent to which the distribution of the publicly provided good is biased towards the more affluent groups relative to the situation of equal shares; i.e., $\gamma$ equal to zero and, thus, $\alpha_i$ equal to $c/m$, where $m$ is the total number of recipients of the publicly provided good.

[10] A transparent illustration of this case is provided by the Cobb-Douglas utility function, $U_i = A z_i^a q_i^{(1-a)}$. In this case,

$$q_i^d = (1-a)\frac{y_i}{\left(\dfrac{\tau_i p}{\alpha_i}\right)}$$

and $$x_i^d = (1-a)\frac{y_i}{\tau_i p}$$

In short, the demand function for $x$ does not then depend on $\alpha_i$.

[11] See Bergstrom and Goodman, Borcherding and Deacon, Robin Barlow, and Bergstrom for similar uses of these particular functional forms.

[12] This interpretation is implicit in the work of Bergstrom. To see this interpretation, assume (i) that the only source of taxation is a fixed percentage tax on the market value of residential property; and (ii) that all individuals have the same demand function for housing:

$$Q_i^H = \theta p_H^\mu y_i^\beta$$

The tax bill of the $i$th individual, $T_i$, is, hence,

$$T_i \equiv t P_H Q_i^H = t\theta P_H^{1+\mu} y_i^\beta$$

where $t$ is the real tax rate, adjusted for the assessment ratio, and the total community tax bill $T$ is

$$T \equiv \sum_j T_j = t\theta P_H^{1+\mu} \sum_j y_j^\beta$$

Therefore, the tax share of the $i$th individual is

$$\tau_i \equiv \frac{T_i}{T} = \frac{y_i^\beta}{\sum y_j^\beta}$$

An example of a distribution scheme that leads to equation (13) is one in which the distribution of the benefits of certain governmental services is in proportion to the value of an individual's property holdings.[13] Another example of benefit share progressivity can be derived from the analysis of Buchanan (1971). In this article, the attempt to "bribe" wealthier central city residents not to migrate to the suburbs leads to a distribution of governmental services which may be biased towards the rich.

In the context of a simple majority rule voting system, changes in $\gamma$, the degree of benefit share progressivity, generate comparative static results on a market level variable such as $x$ by affecting the median quantity demanded—the equilibrium provision of the publicly provided good.[14] For each voter, the effect of a change in $\gamma$ on his demand for $x$ is given by:

$$\frac{\partial x_i^d}{\partial \gamma} = \frac{\partial x_i^d}{\partial \alpha_i}\frac{\partial \alpha_i}{\partial \gamma} \tag{14}$$

with

$$\frac{\partial x_i^d}{\partial \alpha_i} = -\frac{1}{\alpha_i} x_i^d(1+\rho_2) \gtrless 0 \qquad \text{as } \rho_2 \lesseqgtr -1$$

The second factor in equation (14) is more complex but is given by:

$$\frac{\partial \alpha_i}{\partial \gamma} = \frac{\alpha_i \sum_j (ln\, y_i - ln\, y_j) y_j^\gamma}{\sum_j y_j^\gamma} \tag{15}$$

For a given value of $\gamma$, the sign of this expression depends solely on the location of the $i$th voter in the income distribution.

In certain cases—given particular values of $\gamma$ and properties of the income distribution—the sign of $(d\alpha_i/d\gamma)$ and, hence, $(dx_i^d/d\gamma)$ can be determined. For the present problem, Lemma 1 of Bergstrom implies that

$$\frac{\partial \alpha_i}{\partial \gamma} < 0 \qquad \text{if } y_i < \left(\frac{1}{m}\sum_j y_j^\gamma\right)^{1/\gamma} \tag{16}$$

For certain cases this result applies to a majority of voters including the median income voter.[15] In these cases, the effect of changes in $\gamma$ on the *equilibrium* provision (or production) of the publicly provided good $x$ as determined through the voting process can also be established.

Bergstrom's work can readily be used to show that:

$$\frac{\partial x}{\partial \gamma} \gtrless 0, \qquad \text{for } \rho_2 \lessgtr -1 \tag{17}$$

if $\gamma \geqq 1$ and $y_{median} < y_{mean}$

In words, if the demand for the total provision of the publicly provided good is price inelastic (price elastic), benefit shares are highly progressive, and the median income is less than the mean income, then increases in the degree of benefit share progressivity lead to an increase (a decrease) in the equilibrium provision of the publicly provided good. (See Bergstrom, p. 190.)

Next, if $\gamma$ is less than or equal to zero, an additional result can be derived. Note that the limit,

$$\lim_{\gamma \to 0}\left(\frac{1}{m}\sum_j y_j^\gamma\right)^{1/\gamma}$$

is the geometric mean income. In this case,

[13] Noel Edelson makes this assumption in his analysis of a local government solely financed by a property tax. In this case, if $c=1$, the property is homogeneous and is the same property as assessed for tax purposes (see fn. 12), then $\gamma$ equals $\beta$ and benefits are distributed in the same manner as tax bills—$\alpha_i$ equals $\tau_i$. In this case, the tax price to voter $i$ of consuming an additional unit of the publicly provided good $p_i^q$ is equal to the market price of an additional production unit, $p$. In addition, if the effect of the public sector is only on property values and not directly on utility at all, then Edelson has shown that the property tax leads to Lindahl equilibrium and is Pareto optimal (see Section II on Lindahl equilibrium).

[14] See Howard Bowen and Duncan Black for a presentation of the median voter analysis.

[15] In the present model, the ordinal positions of income receivers is either the same or reversed in the distribution of demand quantities. In either case, the median income recipient is the median voter, which will be the equilibrium in a majority voting procedure (see Bowen or Black (ch. 1)).

equations (14) and (16) imply that:

$$\frac{\partial x}{\partial \gamma} \lessgtr 0 \quad \text{for } \rho_2 \gtrless -1 \tag{18}$$

if $\gamma \leqq 0$ and $y_{geometric\ mean} \leqq y_{median}$

In words, if the demand for $x$ is price inelastic (price elastic), benefit shares are equal or more regressive, and the geometric mean income is less than or equal to the median income, then increases in the degree of benefit share progressivity lead to a decrease (an increase) in the equilibrium provision of the publicly provided good.[16]

## II. Applications of the Model

The model developed in Section I can be illustrated by applying it to an analysis of: 1) the public provision of a purely private good, bread; and 2) the public provision of fire services. These two examples illustrate several possible interpretations of the benefit shares, and also demonstrate the importance of studying differences in the benefit shares.

### A. *The Public Provision of a Purely Private Good: Bread*

Consider the following simple example. Suppose a government is publicly providing bread to the voters in a community. McTavish (a voter) is initially in equilibrium at his initial benefit share. Due to some change in policy, his benefit share is doubled. The cost to McTavish of the government providing another loaf of bread to the community (distributed according to benefit shares) is unchanged and still equals his tax share times the market price of a loaf of bread. But, the cost to McTavish of receiving another loaf of bread from the government for his own consumption has been halved. Assuming that McTavish has price-inelastic demand, he will increase his demand for loaves (in consumption units) but to less than twice as much. He will, however, if the total provision remains unchanged, get twice as much bread each week, an amount greater than his demand. Therefore, he will prefer to reduce the size of the public sector, reducing his tax bill.

The above analysis applies only when the private market for bread has been abolished. If an individual can buy or sell bread in the private market, then his voting behavior with respect to the public provision of bread is dramatically affected. Suppose first that an individual can buy bread in the private market to supplement his public allocation, but that resale is prohibited. A voter faced with a consumption tax price $p_i^q$ greater than the market price (i.e, $\tau_i > \alpha_i$) will wish to buy his bread in the market and express a zero demand for publicly provided bread. It is clear then that the public provision of bread at a nonzero level will occur only if a majority of voters have tax shares no bigger than their benefit shares.

Next consider the other extreme in which the public provision cannot be supplemented by purchases in the private market, but the individual can adjust his public allocation by reselling the bread to the government at a price of $p$ (or by selling it in the export market). Under these conditions voters with a tax price for a consumption unit $p_i^q$, below the market price $p$, will vote for unbounded increases in the public provision of bread as they can resell the excess beyond their own consumption to the government, making a per unit profit of $(p - p_i^q)$. For equilibrium to exist, a majority of the voters must have tax prices $p_i^q$ greater than or equal to the market price $p$ (or $\tau_i \geq \alpha_i$).

Combining these results leads to the conclusion that for the public provision of a purely private good to continue in the presence of a private, competitive market, allowing for private purchases and resales,

[16] For a sample of Kentucky counties using 1970 Census of Population data, it turns out that the median income is greater than the geometric mean income for each county.

a majority of voters must have tax prices $p_i^q$ of exactly $p$. In short, continued public provision of a freely exchangeable private commodity requires either: 1) outlawing private exchange of the good; 2) a special earmarked tax with tax shares set equal to benefit shares for a majority of voters; or 3) the setting of benefit shares so that they equal the marginal tax shares implicit in the given finance system. Since the occurrence of either of the last two conditions seems improbable, and the criminalization of private markets generates many other problems, it appears unlikely that easily exchangeable goods such as bread could ever be publicly provided in a participatory democracy.[17] This example also leads to one further observation. Those private commodities that it is feasible to provide publicly are likely to be the same commodities that would allow entrepreneurs to practice price discrimination—services such as medical care and fire protection are obvious examples.

### B. *The Public Provision of Fire Services*

Suppose that the governing body of a town is considering the construction of a firehouse of a certain size. Ignoring the road network, a voter in this town would generally perceive himself to be receiving a larger quantity of fire services if the station is closer to his residence. In terms of the notation of the present paper, the $i$th voter's benefit share $\alpha_i$ is negatively related to the distance of his residence from the firehouse.[18] In these circumstances, if $x$ represents the size of the firehouse, then $\alpha_i x$ is the $i$th voter's perception of the quantity of fire services available to him. Finally, assume that the voter only cares about the fire services provided directly to him.[19]

The application of the present model to this example involves a particular restriction on preferences. Let $d_i$ be the distance that voter $i$'s home is from the firehouse of size $x$. If the voter's preferences are assumed to be representable by

$$U_i = U_i(z_i, x, d_i) \tag{19}$$

then, for the present formulation to be correct, there must exist a function $e(d)$, such that

$$\alpha_i = e(d_i), \qquad e' < 0 \tag{20}$$

and

$$U_i(z_i, x, d_i) = V_i(z_i, \alpha_i x) \tag{21}$$

That is, preferences must be separable as between $(z_i)$ and $(x, d_i)$; i.e., the marginal rate of substitution ($MRS$) between $x$ and $d_i$ is independent of the value of $z_i$. If preferences do not exhibit this separability, then no amount of effort applied to the analysis of the "technological" relation, equation (2), will yield correct results. The distinction between units in production and units in consumption leads to some useful results, but this distinction is not primary. Rather its validity depends upon the form of consumer preferences.

The analysis based on equations (7) and (9) now applies. If the voter has price inelastic demand for fire services, then as the proposed fire station is moved closer to his house, he will desire and vote for a smaller size fire station since his benefit share increases from equation (20). This result does not seem at all unusual, as the closer fire station provides him with a greater amount of fire protection, reducing its marginal value to the voter (his $MRS$ diminishes), hence causing him to wish to

[17] See Gordon Tullock on the likely results of such criminalization.

[18] The interpretation one attaches to $\alpha$ and their resulting properties vary with the problem being examined. Note that in the earlier example of Section II in which bread, a purely private good, is publicly provided, the benefit shares are the actual proportion of the total supply received by each voter and must sum to one. In this firehouse example, no such restriction is implied.

[19] Actually, $\alpha_i x$ could represent the level of fire services for an entire neighborhood, as perceived by the $i$th voter.

substitute lower taxes for fire station size.

Benefit taxation is often suggested for the financing of public goods. In particular, Lindahl tax prices have been analyzed as they lead to both optimality in the level of provision as well as unanimous agreement as to this level. Both results hold if one can find a vector of tax prices, $(p_1^x, \ldots, p_m^x)$, such that the quantity of $x$ demanded is the same for all voters, since on the demand function, tax price is equated to the *MRS* for the public good. Thus, the distribution of Lindahl tax prices will be the same as that of the *MRS*.

If voters have similar tastes and price-inelastic demand for fire services, then the Lindahl tax prices will be *lower* for these consumers closer to the firehouse and *higher* for those further away (given that they have the same income). Thus voters who receive more fire services pay smaller tax bills. This result is quite different from the tax systems used in highway or sewer finance in which tax shares are related positively to benefit shares. The differences follow from the fact that the Lindahl system ignores total benefits and instead deals only with marginal quantities.

## REFERENCES


**B. Barlow**, "Efficiency Aspects of Local School Finance," *J. Polit. Econ.*, Sept./Oct. 1970, *78*, 1028–40.

**T. C. Bergstrom**, "A Note on Efficient Taxation," *J. Polit. Econ.*, Jan./Feb. 1973, *81*, 187–91.

——— **and R. Goodman**, "Private Demand for Public Goods," *Amer. Econ. Rev.*, June 1973, *63*, 280–96.

**D. Black**, *The Theory of Committees and Elections*, Cambridge 1958.

**T. D. Borcherding and R. T. Deacon**, "The Demand for the Services of Non-Federal Governments," *Amer. Econ. Rev.*, Dec. 1972, *62*, 891–906.

**H. R. Bowen**, "The Interpretation of Voting in the Allocation of Economic Resources," *Quart. J. Econ.*, Nov. 1943, *58*, 27–48.

**A. Breton**, "A Theory of Government Grants," *Can. J. Econ.*, May 1965, *31*, 175–87.

**J. M. Buchanan**, "Joint Supply, Externality, and Optimality," *Economica*, Nov. 1966, *33*, 404–15.

———, "Breton and Weldon on Public Goods," *Can. J. Econ.*, Feb. 1967, *33*, 111–15.

———, *The Demand and Supply of Public Goods*, New York 1968.

———, "Principles of Urban Fiscal Strategy," *Publ. Choice*, Fall 1971, *11*, 1–16.

**N. M. Edelson**, "Voting Equilibria With Market-Based Assessments," The Fels Center of Government, disc. pap. no. 48, Univ. Pennsylvania 1974.

**C. R. Plott and R. A. Meyer**, "The Technology of Public Goods Externalities and the Exclusion Principle," soc. sci. work. pap. no. 15 (rev.), Calif. Inst. Techn. 1973.

**R. M. Spann**, "Public vs. Private Provision of Governmental Services," in T. D. Borcherding, ed., *Budgets and Bureaucrats: The Origins of Government Growth*, Durham (forthcoming).

**G. Tullock**, "Subsidized Housing in a Competitive Market: Comment," *Amer. Econ. Rev.*, Mar. 1971, *61*, 218–19.

# The Non-Uniqueness of Equilibrium in the Löschian Location Model

*By* B. Curtis Eaton and Richard G. Lipsey*

The great emphasis placed on regular hexagonal market networks in location theory suggests a presumption that: 1) this network is the unique equilibrium configuration when the number of firms in the market is given, and 2) free entry will cause this network to prevail. These presumptions have been prevalent in the literature since the monumental work of August Lösch. In this paper we argue that both presumptions are wrong: the regular hexagonal lattice is only one of a large number of equilibrium configurations, and free entry does not necessarily produce regular hexagons.

It is well known that (subject to the condition that every point in the space be served) the hexagonal network of market boundaries would be the planner's solution.[1] This result is not at issue here. The question that we consider is whether or not free competition, with each firm seeking to maximize its own profits, will bring about the hexagonal configuration of firms. If this question is to be studied satisfactorily, it is important not to assume the answer at the outset, but rather to make behavioral assumptions suitable to the decisions of independent firms and households and then to determine whether or not the hexagonal configuration will result from the assumed behavior.

In Section I of this paper we analyze a simple model designed to deal with this question. In Section II we relax some of the model's more restrictive assumptions. In the concluding section we discuss the reasons why our results differ from those that appear to follow from Lösch's famous treatment of the same problem.

## I. The Analysis of the Simple Model

### A. *The Model*

We begin by listing and discussing the assumptions of our model.

(a) The infinitely extensible plane is uniformly populated with customers with a density of one per unit of area.

(b) Each firm is faced with the same total cost function, $TC=K+cQ$, where $K$ is fixed costs, $c$ is the constant marginal cost of production, and $Q$ is quantity produced and sold. This cost function is commonly used in location theory and we can take $c$ as zero without loss of generality.

One way in which such a cost function could arise is through indivisibilities. To rationalize this precise function in this way we can assume that the only fixed costs are those associated with capital, and that there is an indivisibility in plant size such that the smallest possible plant is large enough to serve any of the markets for individual firms that we consider. We thus have $K=rI$ where $r$ is the opportunity cost of capital and $I$ is the investment associated with the minimum possible size of plant. Since indivisibilities are ubiquitous with any production involving machines,

* Associate professor of economics, University of British Columbia, and professor of economics, Queen's University, respectively. The present paper is a revised version of the Queen's Institute for Economic Research study paper no. 88, September 1972. We wish to acknowledge the helpful comments and suggestions of many colleagues, the valuable research assistance of Edward France, Allan Popoff, and Duncan Smeaton, and the generous support of the Queen's Institute for Economic Research.

[1] See Belá Bollobás and Nicholas Stern for a recent example of a full discussion of this proposition.

our cost assumption is not empirically uninteresting.[2]

(c) Transport costs per unit distance are constant. Consumers bear the cost of transport.

(d) Consumers buy from the firm whose delivered price is lowest.

(e) All firms charge the same mill price which is an exogenously imposed constant (the common mill price must exceed $c$ and it is taken to be unity).

Assumptions (d) and (e) together imply that the market boundary between two adjacent firms is the perpendicular bisector of the line joining the two firms. Thus each firm's market is defined by a set of linear boundary segments whose positions depend solely on the locations of the firm and its neighbors. Assumption (e) allows us to concentrate on spatial rather than on price competition, but it is not necessary for the results that we obtain. The consequences of dropping it are considered in Section II.[3]

(f) All customers buy one unit of the product per period of time. This assumption greatly simplifies the analysis and exposition, but it is not necessary for the results that we obtain. The consequences of dropping it are also discussed in Section II.

(g) In selecting its location, each firm seeks to maximize its profits and takes the location of all other firms as given, i.e., it has a zero conjectural variation ($ZCV$). With respect to location $ZCV$ is an appropriate assumption in at least two important cases: 1) where equilibrium is approached very rapidly so that firms do not have time to learn their opponents' reactions; and 2) when relocation occurs (if at all) only after a very long time lag, as in many locational problems.[4] (Since our firms are engaged in noncooperative games, any equilibria that we discover will be Nash equilibria.)

The assumptions made above affect the measurement of our profit variables. *Gross profit* $R$ is total revenue minus total variable cost: $R=PQ-cQ$. Since we have assumed $c=0$ and $P=1$ and, since each unit of market area contains one customer who buys one unit of the product per period of time, we have $R=Q=$ *Market Area. Pure profits* $Z$ (also called net profits) are gross profits minus total fixed costs: $Z=R-K$. Thus in the model of this section, maximizing profits is the same thing as maximizing market area.

[2] As Nicholas Kaldor puts it, "Yet on an empirical level, nobody doubts that in any economic activity which involves processing or transforming basic materials—in other words, in industry—increasing returns dominate the picture for the very reasons given by Adam Smith in the first chapter of *The Wealth of Nations:* reasons that are fundamental to the nature of technological processes and not to any particular technology" (p. 1242). See also Paul Samuelson. Indeed if there were no indivisibilities and genuine constant returns to scale everywhere in a spatial economy in which transport costs were positive, then there would be no firms at all since all production would take place at the point of consumption so as to reduce transport costs to zero.

[3] It is important to note that our model with a parametric price that is common to all firms will have comparative static properties that are qualitatively similar to those of other models of price competition in which there is a single equilibrium price common to all firms. Consider, for example, the comparative static properties that are implied by the price-conjectural variation that is implicit in Edwin Mills and Michael Lav, and explicit in Martin Beckmann (1971). These authors assume that when setting its price, the firm assumes that each of its competitors will adopt the same price. With this assumption, as with our parametric price, the firm's anticipated market area is independent of the firm's own price. In the Mills-Lav and Beckmann case the existing firms charge the common profit-maximizing price and the general level of profits is raised or lowered by exit or entry. In our model the firms charge a common arbitrary price and the general level of profits is raised or lowered by exit or entry. Indeed anything that we prove for any arbitrary price must be true for one particular price, the monopoly price. Qualitatively these models behave in a similar fashion with respect to spatial competition.

[4] This does not mean that relocation never occurs, but only that the firm's locational decision is taken on the assumption that *over the relevant planning horizon* other firms will not relocate in response. Of course other conjectural variations can be studied, but in each new case the conditions of equilibrium will have to be suitably amended.

These assumptions outline the simplest model that will do our job. We wish to consider the conjectures: (i) that the Löschian hexagonal pattern is the unique equilibrium configuration on the infinitely extensible plane, and (ii) that if other configurations are arbitrarily imposed, the process of entry of new firms will convert these into hexagonal configurations. Our model provides a counterexample which is sufficient to refute both of these conjectures. It also shows some other configurations that can sometimes produce equilibria and it allows us to examine the relative stability of each of these under conditions of entry. In Section II we relax some of the assumptions of this model in order to deal with the possible objection that our assumptions are so restrictive that our counterexamples are of no real interest other than as pathological cases.

### B. *The Technique of Analysis*

We analyze the behavior of our model using numerical simulation techniques. We do this because most of the questions we wish to ask are extraordinarily difficult, if not impossible, to answer using conventional analytical methods.

Where they are feasible, analytical techniques are usually preferable to numerical simulation techniques. Numerical methods can prove, for example, that a given configuration *is not* an equilibrium configuration by demonstrating the existence of a profitable move. They cannot, however, prove that a particular configuration *is* an equilibrium configuration. They are only strongly suggestive. The inevitable rounding errors of simulation routines limit the fineness of the grid of potential locations, and in any case it is impossible to evaluate potential profits at an infinite number of points. We have used numerical techniques solely because the problem seemed intractable with conventional analytical techniques, while numerical techniques do seem to allow us to proceed further than other investigators who have confined themselves to analytical methods.

The core of our technique is an algorithm (described in detail in our 1972a paper), which answers the following question: given the location of $n-1$ firms in some region of the infinitely extensible plane, what market can the $n$th firm expect to control if it locates at an arbitrary point in the region?[5] Using this algorithm it is an easy matter to produce a map which describes the profits that the $n$th firm could expect to have if it located at any one of a large number of alternative points, given the locations of all of its neighbors. We refer to such maps as market maps.

Figure 1 reproduces a part of one of these maps. The origin is shown by the dot within the circle, and the dots are spaced four units apart vertically and horizontally. The neighbors of the $n$th firm are shown by circled crosses. The numbers without parentheses indicate the $n$th firm's market area and hence by the assumptions in Section I, its gross profits in each of the positions indicated by the dots. Thus, for example, if the firm locates at the origin, its market area, and hence its profits, are 692 while if it locates at the point (8, 4) they are 698. (The numbers in parentheses show the profits in each loca-

[5] The algorithm was developed to analyze locational problems in a bounded market (see the authors, 1975a), but is easily adapted to the study of some types of behavior in infinitely extensible Löschian space. The basic analytical difference between unbounded and bounded space is that in the former a firm's market boundary must be composed entirely of boundary segments with other firms, whereas in bounded space a firm's market boundary can coincide over some of its range with the boundary of the market. Consider the problem of determining a firm's market area at some particular point in infinite space. The firm will, at least in all the cases in which we are interested, have a boundary with a small number of neighboring firms. The market area algorithm for the bounded space is adapted to the problem of a firm locating at a particular point in unbounded space simply by including all of the neighboring firms in the bounded space.

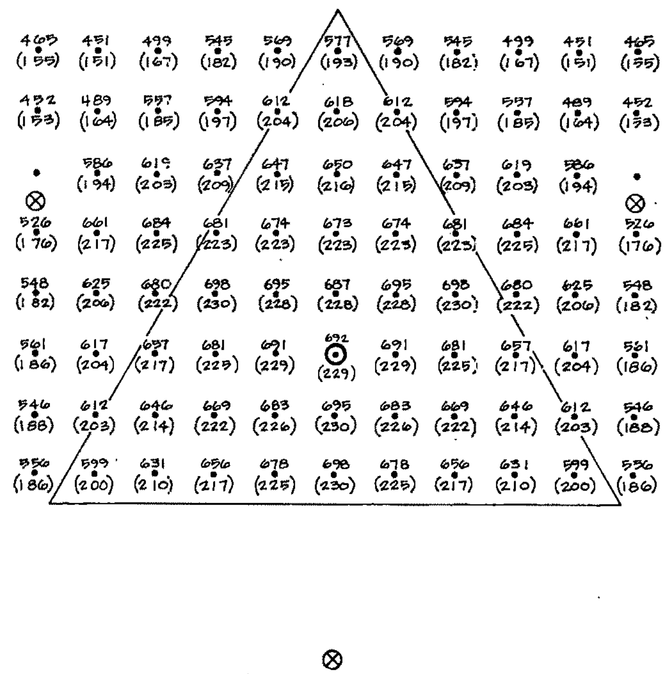



FIGURE 1. THE ABSENCE OF EQUILIBRIUM IN A NETWORK OF EQUILATERAL TRIANGLES

The numbers show the profits the $n$th firm would make by locating alternatively at each of the dots and with neighbors in the fixed positions shown by the circled crosses. Numbers in parentheses refer to the model of Section II, numbers not in parentheses refer to the model of Section I. Location of the $n$th firm at the origin, the point inside the circle, completes the symmetrical lattice of firms and gives the $n$th firm the market boundaries indicated by the dashed lines. Rows and columns are separated by 4 units in Figure 1 and by 6 units in all other figures.

tion that arise from the model in Section II.)

In our actual program, profits were calculated for many more points than are shown in our figures. Only 1/16 of the points are shown in Figure 1, and 1/36 of the points in all other figures. Furthermore, if we were uncertain about the precise location of a particular equilibrium, we could magnify any particular part of the market by placing our observation points as close together as we wished.

### C. *Equilibrium in the Absence of Entry and Exit*

We now pass to a consideration of the behavior of our model. We consider, first, a situation in which firms are allowed neither to enter nor to leave the market. In the absence of entry and exit there is a single condition that is necessary and sufficient for all firms in the market to be in equilibrium:

Equilibrium Condition (i): *No firm can find a new location that offers it a larger anticipated profit (assuming ZCV) than that obtained in its present location.*

We first ask which of the three configurations of regular, space-filling polygons (triangles, squares, and hexagons) satisfy equilibrium condition (i). It is interesting that no answer to this question appears to exist in the literature. The reason probably lies with the extreme diffi-

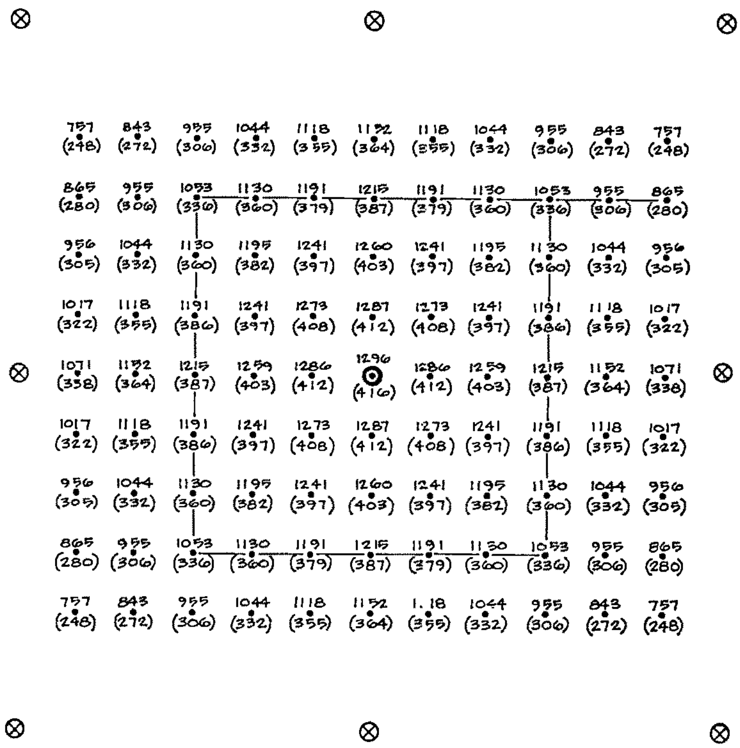



FIGURE 2. EQUILIBRIUM IN A NETWORK OF SQUARES
For the description, see Figure 1.

culty of using conventional analytical methods to deal with such problems.

To answer this question using our technique, we let all the firms be first arranged so that the network of market boundaries is composed of equilateral triangles. We select one firm and make its location the origin. We allow that firm to consider a large number of alternative locations and calculate the market area and hence the profits that it would obtain in each location. The triangular configuration is an equilibrium configuration if and only if the best location for the firm is at the origin. Next, repeat the experiment with the firms arranged so that there is a square network of market boundaries and, finally, so that there is a regular hexagonal network of market boundaries.

Figures 1, 2, and 3 show the results. Intuition suggests and calculation confirms that the firm is always better off remaining within the area defined by its present neighbors rather than moving outside of that area to relocate in an already complete portion of the lattice of firms. Our maps are thus confined to this area. The broken lines indicate the firm's market boundaries when it locates at the origin, thus completing the regular lattice of firms.

Figure 1 shows that the triangular configuration does not satisfy condition (i): the firm at the origin, and hence *any* existing firm, wishes to relocate. The firm would prefer to locate at any of the three locations $(-8, 4)$, $(8, 4)$, or $(0, -8)$ where profits are 698 rather than at the origin where profits are only 692. Thus, given a chance, any firm will move from its posi-


FIGURE 3. EQUILIBRIUM IN A NETWORK OF REGULAR HEXAGONS

For the description, see Figure 1.

tion in a regular triangular lattice. Figures 2 and 3 show that both the square and hexagonal configurations appear to satisfy condition (i): the firm at the origin, and hence any existing firm, does not wish to relocate since the market area and profits at the origin are higher than those at any other location. Thus two of the possible configurations of regular, space-filling polygons satisfy condition (i).

So far we have followed established practice in considering only those configurations of firms that give rise to identical, regular, space-filling polygons. We easily found counterexamples, however, to the conjecture that the only equilibrium configurations were regular, space-filling polygons. We found, for example, that some configurations of identical rectangles would fulfill equilibrium condition (i). Specifically, we discovered by numerical experimentation that condition (i) is fulfilled by any rectangular lattice of market boundaries in which the ratio of the long to the short side of the rectangle is 26:10 or less. Figure 4 provides an example in which the ratio of the sides is 2:1. The figure shows that a firm that is surrounded by a rectangular lattice of other firms maximizes its market area and its profits (at a value of 968 in the present example) by locating at the origin, thus completing the rectangular lattice.

A network of firms and market boundaries composed of irregular hexagons also fulfills condition (i). In Figure 5 the firm that is free to move chooses to locate at the origin, earning gross profits of 1224 and completing the lattice of irregular hexagons. When it does this, each firm is

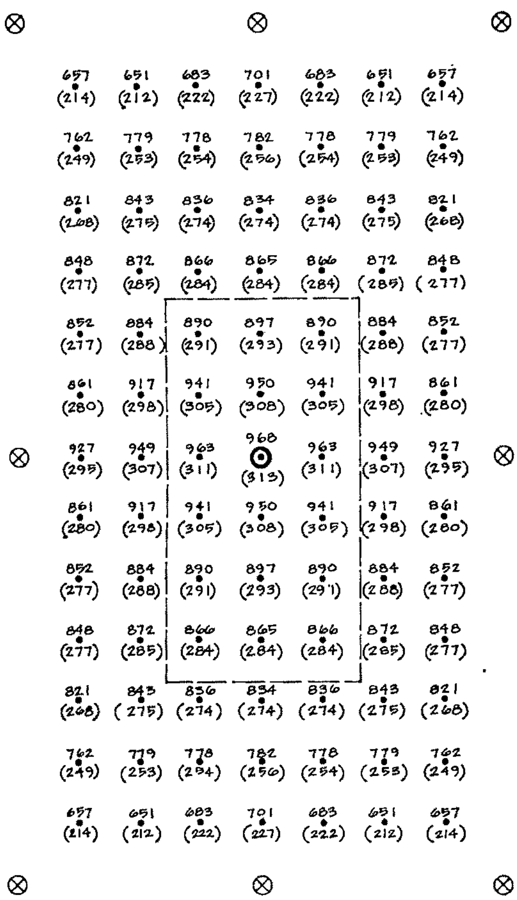



FIGURE 4. EQUILIBRIUM IN A NETWORK OF RECTANGLES WHOSE SIDES STAND IN THE RATIO 2:1

For the description, see Figure 1.

separated from two of its neighbors by 35 and from the remaining four neighbors by 39. This result shows that equilibrium condition (i) can be fulfilled by some configurations that give the firms identical, but nonregular, hexagonal market boundaries.

Our results suggest the further questions: "Can condition (i) be fulfilled: (a) if firms have markets that are not identical in shape but which are equal in area, and (b) if firms do not even have equal market areas?" Although we have not been able to show that identical market shapes or even equal market areas are required by our assumptions, neither have we yet succeeded in finding a configuration of nonidentical markets that satisfies condition (i). *In everything that follows, the argument is confined to situations in which the equilibrium configuration gives all firms markets that are identical in size and shape.*

### D. *Equilibrium Configurations Under Freedom of Entry and Exit*

We now consider the case in which firms are permitted to enter and to leave the market. Condition (i) remains an equilibrium condition, but there are now two further conditions.

Condition (ii): *All possible locations for a new entrant within the network of existing firms offer anticipated gross profits, $R^e$, of less than K.*

Condition (iii): *No existing firm earns actual gross profits, $R^a$, of less than K.*

Taken together, (i), (ii), and (iii) are necessary and sufficient conditions for equilibrium in our free-entry model: (i) ensures that no existing firm wishes to relocate elsewhere in the market; (ii) ensures that no new firm wishes to enter; and (iii) ensures that no existing firm wishes to exit.

We now ask: Which of the configurations that satisfy condition (i) will also satisfy conditions (ii) and (iii)? Our measure of density of packing $h$ is the reciprocal of the distance from a firm to its nearest neighbor. Within any configuration which gives all firms identical markets this definition of density of packing involves no ambiguity.[6]

[6] This definition of density of packing is sufficient for our purposes. In configurations composed of any one of the three regular space-filling polygons the distances to all neighbors are the same. In the irregular configurations that we consider, networks of irregular hexagon and the rectangles, there is no ambiguity in the measure since we hold constant the relative distances from a

FIGURE 5. EQUILIBRIUM IN A NETWORK OF IRREGULAR HEXAGONS
For the description, see Figure 1.

We consider a particular configuration that gives all firms identical markets, which means that the ratios of the distances from a firm to each of its neighbors is held constant. We then vary the density of packing thus changing the absolute distances from each firm to each of its neighbors. The market area, and hence the gross profit of each existing firm, is a monotonically decreasing function of $h$: $R^a = R^a(h)$, with the properties

$$\frac{dR^a}{dh} < 0, \qquad \lim_{h\to 0} R^a = \infty$$

and $\lim_{h\to\infty} R^a = 0$

Now let $R^e$ be the anticipated gross profits of a potential new entrant in its best possible location. Since the profits that a new firm can expect to earn (on a *ZCV* assumption) can be expressed as a function of the profits earned by existing firms which are in turn a function of $h$, we express $R^e$ also as a function of $h$: $R^e = R^e(h)$.

Condition (ii) is fulfilled whenever $R^e(h) < K$. Condition (iii) is fulfilled whenever $R^a(h) \geq K$. Any new entrant must fit into an already completed lattice of firms and must expect to earn profits that are substantially lower than those earned by existing firms before entry occurs. Thus for any given value of $h$, $h_0$, $R^e(h_0) < R^a(h_0)$. To illustrate, numerical calculations show the ratios for $R^e/R^a$ to be approximately 0.51 for a regular hexagonal configuration of firms, 0.56 for a square configuration, and 0.50 for a configuration of rectangles

firm to each of its neighbors. We hold the relative distances constant because we are only interested in whether or not a configuration can be made to satisfy the conditions for entry equilibrium by changing the density of packing. For a general, rigorous treatment of density of firms per unit of market area see Bollobás and Stern.

where the ratio of the lengths of the market boundaries is 2.0. (Although $R^e$ and $R^a$ are both functions of $h$ their ratio is independent of $h$ in each of the configurations that we consider.) It follows immediately that for any configuration of firms that satisfies condition (i), we can find a *range* of values for $h$ that will allow conditions (ii) and (iii) to be fulfilled simultaneously.[7]

Thus the answer to the question with which we began this section is simply that any of the configurations that give identical market areas and that satisfy condition (i) can be made to satisfy (ii) and (iii) by packing the firms closely enough together so that the expected gross profit of any new entrant is less than $K$ and far enough apart so that the actual gross profit of any existing firm is $K$ or more.

It follows that equilibrium in our free-entry model is consistent with a multiplicity of configurations of firms (squares, rectangles, irregular and regular hexagons), and with a *range* of density of packing of firms (and hence of profits for each firm) in *each* of these configurations. The second of these conclusions has already been stated by Beckmann for hexagons (1968, p. 44). The first conclusion does not seem to have been stated in the literature before. We wish to emphasize that the multiplicity of equilibrium configurations would continue to exist even if zero profits were made a condition of equilibrium. For any of our configurations that satisfy condition (i) there is always a density of packing that will make profits exactly zero, thus satisfying conditions (ii) and (iii) at zero profits.

### E. *The Effect of the Process of Entry on the Configuration of Firms*

We may wonder if the dynamic process of free entry is more likely to produce one of the equilibrium configurations, particularly regular hexagons, rather than any other. To consider this, we focus on two specific questions. Does the process of free entry into a square or rectangular configuration of firms tend to transform it into the regular hexagonal configuration? Does the process of free entry into a regular hexagonal configuration tend to reproduce a regular hexagonal configuration?

Many entry dynamics are possible. The one that we use is based on a dynamic suggested for one-dimensional markets by H. Stephen Grace. We discover the set of points of entry offering the highest possible anticipated profits, on the assumption that there is only one new entrant. We then assume that there is simultaneous entry of one firm at each of these points.

Since we know from the previous section that any configuration in which condition (i) is satisfied for all firms can also be made to satisfy conditions (ii) and (iii) by suitable choice of $h$, we concentrate only on condition (i) in the entry process. This allows us to avoid a lengthy discussion of special cases which obscures the main points of interest.[8]

We first consider entry into a network of rectangular markets (in which lengths of the long and short sides of the market boundaries are $X$ and $Y$). A new entrant's most profitable location is at the midpoint of the short side of an existing firm's mar-

[7] To show that $R^e(h_0)<R^a(h_0)$ when condition (i) is satisfied for all existing firms, we use a proof by contradiction. Assume that condition (i) is satisfied for all existing firms (each of whom is making profits of $R^{a*}$ before a new firm enters the market). We denote the profits that a potential new entrant expects as $R^{e*}$. If $R^{a*}<R^{e*}$, the assumption is obviously contradicted, since the entry position was available to any existing firm. If $R^{a*}=R^{e*}$, the assumption is also contradicted. Let $(x_e, y_e)$ be the new entrants' best location, and $(x_i, y_i)$ be the location of its $i$th neighbor. If the $i$th neighbor had chosen the point $(x_e, y_e)$ before the new firm entered, it would have had profits greater than $R^{e*}$ because it would not have had a neighbor at $(x_i, y_i)$. Thus if $R^{e*}$ were equal to $R^{a*}$, then the $i$th neighbor could have had profits greater than $R^{a*}$ if it had located at $(x_e, y_e)$; this contradicts the assumption that condition (i) is satisfied for the $i$th neighbor at the point $(x_i, y_i)$.

[8] See our (1972b) paper for a full discussion of these lengthy obscurities.

ket boundary.[9] When firms enter at each such point, the number of firms is doubled, and a new configuration of rectangular market areas (with sides of lengths $X/2$ and $Y$) is established. Thus the ratio of the long to the short side is changed by entry, but a new configuration of identical rectangular market areas is established.[10] If the configuration satisfied condition (i) before entry, it must satisfy it after entry.[11] Thus rectangles beget rectangles after each round of entry.

We next consider entry into a square network. To obtain the results shown in Figure 6, firms were placed in a square lattice and a single new entrant was allowed to calculate its market area, and hence its profits, at a large number of alternative entry points. The figure shows that the most profitable location for a new entrant is at the midpoint of each of the existing market boundary segments (where profits are 729). Letting a firm enter at each of these points produces the configuration shown by the circled crosses (original firms) and the dots (new entrants) in Figure 7. Condition (i) is no longer satisfied for the original firms. (The diamond in the top left-hand corner of Figure 7 shows one of the best possible points for relocation of the original firms; and such points recur throughout the market in similar locations.) *If* profits after entry were high enough to encourage a second round of entry and *if* this entry occurred before any relocation of firms already in the market, the square network would then be reestablished. The locations of the firms that would come into the market on the second round of entry are shown by the triangles in Figure 7. Condition (i) is now satisfied for all firms. Thus, squares beget squares but only after two rounds of entry and a fourfold increase in the number of firms.

Finally, we consider entry into a hexagonal network. Because the hexagon does not easily reproduce itself, the entry results are complex and their description becomes tedious. We nonetheless thought it worth briefly describing the process because of the enormous amount of attention that has been paid in the literature to the hexagonal configuration.

Beckmann (1968, p. 44) conjectured that a new entrant would wish to locate at the centroid of the equilateral triangle defined by three contiguous firms. If we let a firm enter at each such point, the number of firms is tripled and all firms again have identical, regular, hexagonal market areas. When we applied our simulation model to the entry problem in a hexagonal network, we discovered that Beckmann's conjecture does not apply in our model.[12] Figure 8 shows the market areas for a new

[9] We have not reproduced the map for the rectangular case since it is so similar to the square case shown in Figure 6. In the square the best entry points bisect all the sides of the existing firms' market boundaries.

[10] The original ratio of the long to the short side of the market is $X:Y$. There are several cases: (a) $Y<X<2Y$. After entry, $Y$ becomes the long side and the new ratio becomes $2Y:X$. After a second round of entry the ratio becomes $2X:2Y$ which is, of course, the original ratio. In the special case in which $X:Y=\sqrt{2}$ the ratio is unchanged after each round of entry since if $X/Y=\sqrt{2}$, then $2Y/X=\sqrt{2}$; (b) $X=2Y$. One round of entry creates a square configuration and after this the analysis of entry into square markets is appropriate. This is the special case in which rectangles do not beget rectangles; (c) $2Y<X$. In this case the short side remains the short side after entry, and the ratio of the long to the short side becomes $X:2Y$. Entry proceeds through $r$ rounds until $X<2^rY$ and after this the analysis of case (a) applies.

[11] a) If $X<2Y$, the ratio of the long to the short sides alternates between two values $X:Y$ and $2Y:X$ which must both lie between the bounds of 1 and 2. b) If $X>2Y$, the ratio declines through successive rounds of entry until it reaches a value of less than 2 after which it alternates as in a) above. Thus, if the ratio of the sides is small enough so that condition (i) is satisfied in the initial configuration, entry can never cause it to increase to the point at which condition (i) is not satisfied.

[12] Beckmann's analysis appears to refer to any downward-sloping linear demand curve. Section II below shows that our results generalize at least to some downward-sloping demand curves and that Beckmann's conjecture cannot therefore be correct in general.

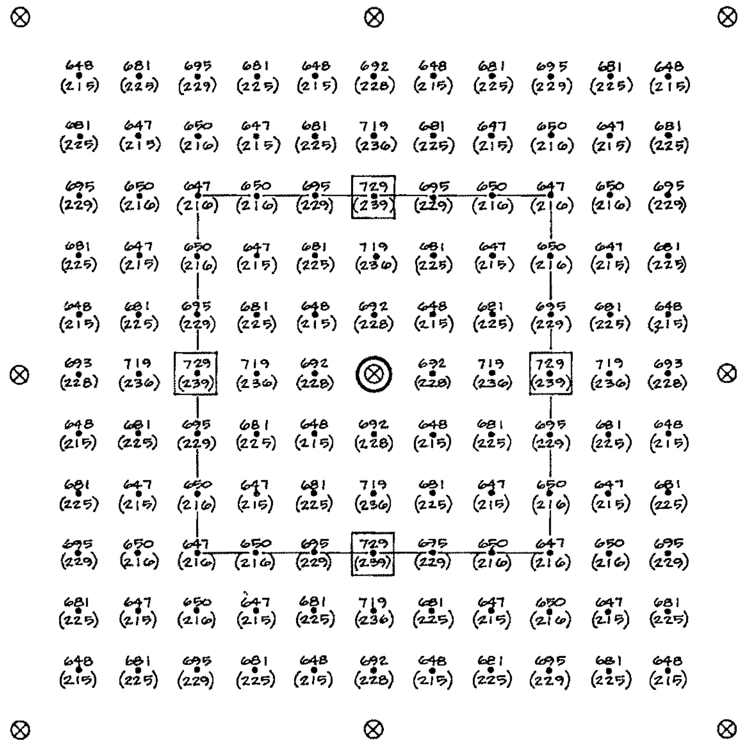



FIGURE 6. ENTRY INTO SQUARE NETWORK

The *n*th firm is located at the origin. Its location is shown by the double circled cross and its market boundaries by the broken lines; its neighbors are shown by single circled crosses. The numbers (in parentheses for the model of Section II and not in parentheses for the model of Section I) show the gross profits available to a single new entrant if it were to locate alternatively at each of the positions indicated by the dots. The best entry points are indicated by squares around both the dot that indicates the exact best point and the numbers that indicate gross profits at that point.

entrant in a number of alternative locations when existing firms are located in a hexagonal configuration at the circled crosses. Beckmann's conjectured entry points are shown by the six points marked with triangles. The market-maximizing entry points turn out to be much closer to the existing firms than Beckmann conjectured. The best entry points close to the firm located at the origin are marked by squares, and there are six such points around each of the existing firms. (The gross profits at these best entry points are approximately 708 while they are 692 at Beckmann's conjectured entry points. Note that because we are only reporting 1/36 of our actual observations our reported observations are close to but not actually at the best points.) If we let a new firm enter at each of the best entry points, we obtain the configuration shown by the circled crosses (original firms) and the dots (new firms) in Figure 9. This is not an equilibrium configuration since condition (i) is no longer satisfied for the original firms. These firms now find themselves surrounded by six very near neighbors and they could substantially increase their market areas by relocating. (The best places to relocate are shown by the triangles in Figure 9.)

In order to carry the analysis a stage

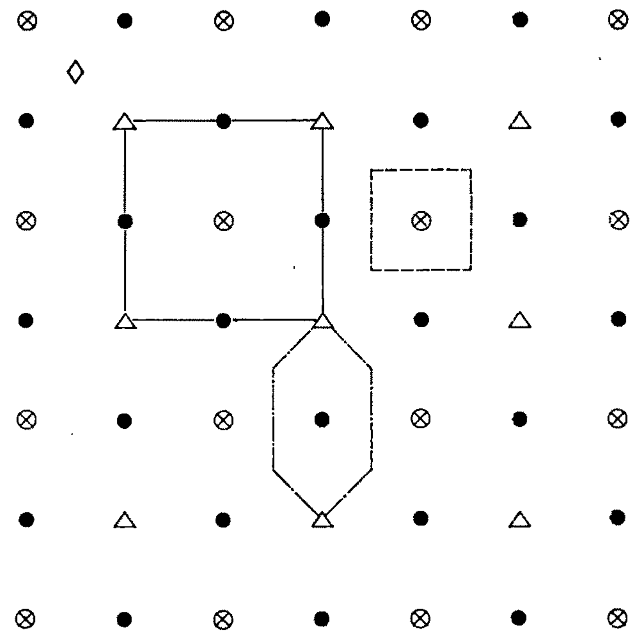


FIGURE 7. ENTRY INTO A SQUARE NETWORK

1) The original firms are shown by circled crosses. The market boundary of one such firm is shown by the solid lines. 2) The locations of new entrants after one round of entry are shown by dots. The market boundary of one such firm is shown by the dot-dashed line. 3) The locations of the new entrants after the second round of entry are shown by the triangles. The market boundary for one such firm is shown by the dashed lines (all firms, old and new, have identical market boundaries after the second round of entry).

further, we consider the effects of further rounds of entry that occur *before* existing firms are allowed to relocate. The second and third rounds of entry are shown in Figure 9 by the triangles and the squares, respectively. The second round still leaves a very irregular network of firms but a third round, if it occurs, reestablishes a hexagonal network. The three rounds, however require a twelvefold increase in the number of firms and, furthermore, the new configuration is not one of regular hexagons. There are no less than four types of market areas. The original firms have regular hexagonal markets, the firms that entered in the three rounds of entry all have irregular hexagonal markets. Condition (i) is not satisfied for the firms that entered in the first and second rounds of entry.

Finally, consider allowing those existing firms for which condition (i) is not fulfilled to relocate. We allow all of the firms that entered in the first round to relocate to their market-maximizing position. The first-round-entry firms are shown with a dot and the desired relocation is shown for six of them by the arrows in Figure 9. Calculation of the market-maximizing locations for these firms shows that this one set of movements establishes a regular hexagonal network of firms and no further movement is desired by any of the firms in the market.

The most important of our entry results may now be summarized. Entry into a square or a rectangular lattice does not tend to turn it into a hexagonal lattice. One round of entry into a hexagonal (or a square) lattice produces disequilibrium

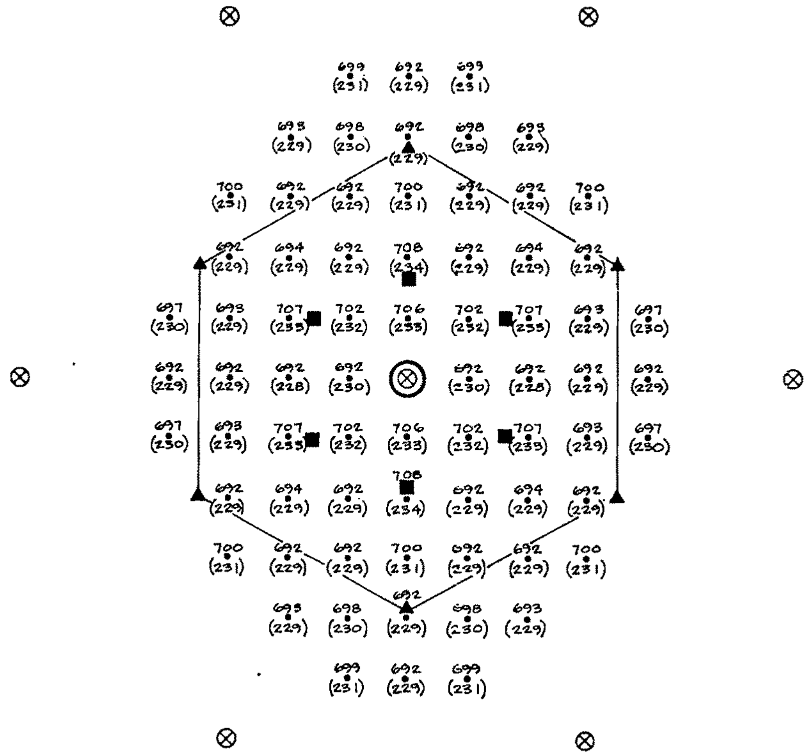


FIGURE 8. ENTRY INTO A REGULAR HEXAGONAL NETWORK

The *n*th firm is located at the origin. Its location is shown by the double circled cross and its market boundaries by the broken lines; its neighbors are shown by single circled crosses. The numbers (in parentheses for the model of Section II and not in parentheses for the model of Section I) show the gross profits available to a new entrant who locates alternatively at each of the positions indicated by the dots. Simultaneous entry at the points indicated by triangles would recreate a hexagonal configuration. The best entry points, however, are those indicated by the squares.

which tends to destroy the lattice. Of the three, therefore, the rectangular lattice seems to be the most robust and the hexagon the least robust in the face of the type of entry that we have considered.

## II. Elasticity of Demand and Price Competition

In this section we relax two of the most restrictive assumptions of our model. In order to focus on the nature of equilibrium configurations in space, we have so far used a model in which the individual consumer's demand is perfectly inelastic (assumption (f)), and we have abstracted from price competition (assumption (e)). These restrictive assumptions considerably simplify the analysis and the exposition of the model but it is important to ask if they are necessary for the results that we obtain. In this section we simultaneously relax both assumptions and demonstrate that our results are not dependent on them.

We now assume that all individuals have the same downward-sloping demand function of a particular form and that firms maximize with respect to price as well as location. The demonstration that our results hold for a particular demand function and a particular type of price competition

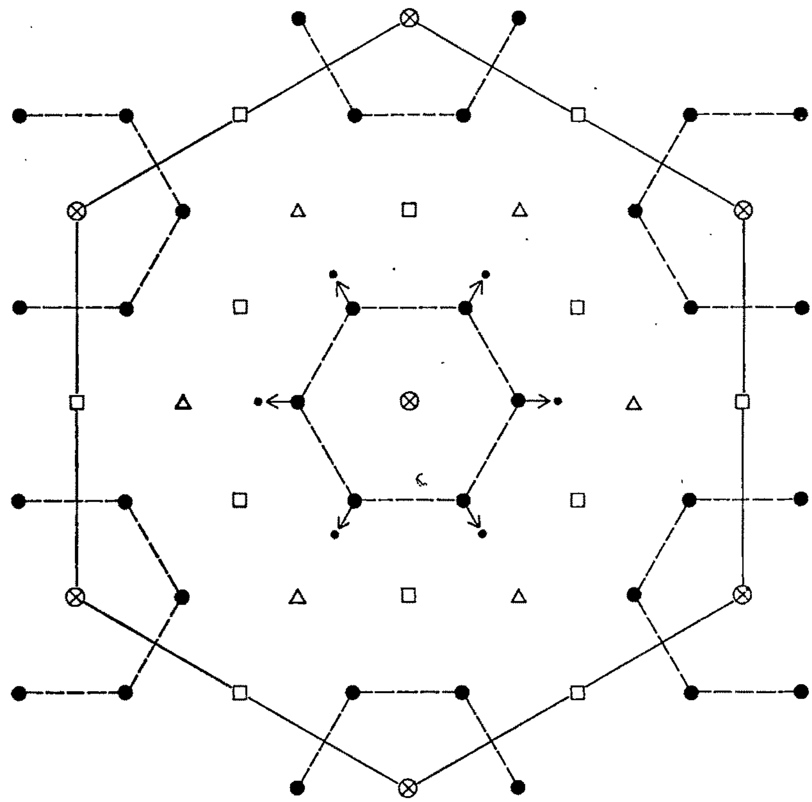


FIGURE 9. ENTRY INTO A HEXAGONAL NETWORK

Original firms are shown by circled crosses. Firms that enter on the first, second, and third rounds of entry are shown by the dots, triangles, and squares, respectively. Market areas are not shown. The lines are solely for visual aid: the original firms surrounding the one at the origin are linked by a solid line; each set of six first-round-entry firms that cluster around each original firm is indicated by a dashed line.

is a counterexample which disproves the conjecture (made by many readers of earlier versions of this paper) that our results are dependent upon our assumptions (e) and (f).

We assume that all customers have identical demand functions of the form

$$q = \exp\left[-P_d\right]$$

where $q$ is quantity demanded and $P_d$ is delivered price. For simplicity, we let transport costs be 1 per unit of distance. Delivered price at any point $(X, Y)$ is then

$$P_d = P + \left[(X - X_0)^2 + (Y - Y_0)^2\right]^{1/2}$$

where $(X_0, Y_0)$ is the location of the firm and $P$ the firm's mill price.

We adopt the conjectural variations with respect to price which is implicit in Mills and Lav and explicit in Beckmann (1971). Any firm $X$ in choosing its price assumes that all neighboring firms will charge the same price that $X$ chooses. The firm maximizes expected profits with respect to price at any given location.

This price-conjectural variation implies that, just as in the model of Section I, the firm's anticipated market boundary will be composed of linear segments which are the perpendicular bisectors of the lines through the firm's own location and the locations of its neighboring firms. The firm's an-

ticipated market area is independent of its price.

The firm's aggregated demand function will be

$$Q = \iint_{MA} \exp\left[-(P + ((X - X_0)^2 + (Y - Y_0)^2)^{1/2}\right]dxdy$$
$$= \exp[-P] \iint_{MA} \exp\left[-((X - X_0)^2 + (Y - Y_0)^2)^{1/2}\right]dxdy$$

The integration is over all points $(X, Y)$ within the firm's anticipated market area, $MA$. This aggregated demand function will have the form

$$Q = M \exp[-P]$$

where $M$ is dependent upon the firm's market area. With zero marginal costs of production, profit maximization at any location requires that the firm find the point of unit elasticity on its aggregated demand function. The point of unit elasticity occurs at a price of 1, and the firm's expectations with respect to its neighbor's price will be correct given that all firms pursue the same profit-maximizing behavior with the same price-conjectural variation.

All other assumptions of Section I are maintained. We analyze the firm's profit-maximizing behavior in this modified model using numerical techniques which are analogous to those described in Section I. At each point, we numerically evaluate the aggregated demand function. We ask the same questions of the revised model as were asked of the model in Section I. The numbers in parentheses in Figures 1–5 show data for this new model analogous to the data already referred to for the model in Section I. Since in this model Market Area $\neq Q \neq TR =$ Gross Profit, we must choose which of these variables to plot. The relevant variable is total revenue, which is the same as gross profits (since Marginal Cost$=0$). In Figure 1, for example, the firm at the origin earns gross profits of 229 in our present model and of 692 in the model of Section I. In both cases, the firm does not wish to locate at the origin (it can earn 230 in the present model and 698 in the model of Section I at the points of maximum profits). In Figures 2 through 5, location at the origin does maximize profits. Since location of the $n$th firm at the origin completes the configuration in each case, we conclude that, just as with the model in Section I, equilateral triangles are not an equilibrium configuration, while configurations of squares and rectangles appear to be equilibrium configurations, as do configurations of regular and irregular hexagons.

Next, we consider entry (on the assumption that profits are large enough to induce it). Using the entry dynamic described in Section I, we discover that rectangles beget rectangles in one round of entry. In both the square and regular hexagonal networks (see Figures 6 and 8) one round of entry does not reproduce the configuration and there is therefore a tendency for these configurations to break up. These results are the same as those for the first round of entry in the model of Section I.[13]

### III. Conclusions

(1) There is a wide range of configurations of firms—including squares, rectangles, regular and irregular hexagons, but not equilateral triangles—that satisfy the equilibrium condition of our models

[13] In subsequent work we have discovered that the best entry point varies with the elasticity of demand. Our results apply as long as elasticity is less than some critical value which varies with the parameters of the model. Above one critical value, rectangles do not reproduce themselves, while above another (very large) critical value, hexagons do reproduce themselves. These relationships are the subject of further research and the results will be published in study papers from the University of British Columbia and Queen's University under the title "Entry Experiments in Löschian Space."

with respect to the location of firms already in the market.

(2) Any of the configurations that satisfy the equilibrium condition for the location of firms already in the market can be made to satisfy the equilibrium conditions for free entry by packing the existing firms densely enough so that the expected profits of a new entrant are negative and loosely enough so that the profits of existing firms are nonnegative.

(3) Zero profits is not a condition of entry equilibrium in our model and thus there is a wide range of density of packing of firms in each of the configurations mentioned in conclusion (1) above that is consistent with entry equilibrium.

(4) On the dynamic assumption about entry that we have investigated, there is no tendency for entry to convert a nonhexagonal configuration into a hexagonal one; indeed a rectangular configuration is more likely to persist through rounds of entry of new firms than is a hexagonal configuration.

(5) The model used to establish conclusions (1), (2), (3), and (4) assumes zero elasticity of demand and an exogenously imposed parametric price. In Section II we demonstrated that none of these conclusions is critically dependent upon these two assumptions. In that section each customer has a downward-sloping demand curve and firms maximize with respect to both price and location. Conclusions (1) through (4) continue to hold.

Given that our results differ from those in the existing literature, we must ask why these differences arise. The basic reasons seem to be methodological.[14] Lösch and many of the location theorists who followed him did not make explicit conjectural-variation assumptions concerning either or both location and price (see, for example, Lösch, pp. 94–97). It is impossible, however, to know what profits a firm expects to earn in alternative locations unless we know what reaction it expects its neighbors to make to changes in its location and price. Therefore, in the absence of conjectural-variation assumptions, equilibrium is undefined and equilibrium conditions cannot be derived.

Not having well-defined equilibrium conditions, many location theorists beginning with Lösch (see pp. 94–97) have imposed what to them seemed to be reasonable equilibrium conditions. We may consider by way of illustration two of the commonly assumed conditions of zero profits and densest packing.[15]

The zero-profits condition is imposed on the argument that as long as the profits of existing firms are positive, it will pay new firms to enter. But this argument assumes that the expected profits of a new entrant, $R^e$, are the same as the actual profits of a typical existing firm, $R^a$ (any existing firm will do as long as they have identical market areas). But Beckmann has shown (1968, p. 44) that there is a class of models for which $R^e(h_0) < R^a(h_0)$ so that for $R^e(h_0) = 0$ it is not necessary that $R^a(h_0) = 0$ (where $h_0$ is any specific density of packing).[16]

The assumed equilibrium condition of densest packing gives rise to the assumption that a regular hexagonal lattice is the

[14] See our (1972b) paper for a much more detailed discussion of this problem than is given here.

[15] For examples two recent articles, Kenneth Denike and John Parr, and John Hartwick, impose both zero profits and densest packing as conditions of equilibrium without deriving them from behavioral assumptions.

[16] Another commonly used argument is that if existing firms are making positive pure profits, a new firm could enter right alongside the old firm, slightly undercut its price, and drive the old firm out of business getting *all* of its profits. If firms make a *ZCV* assumption with respect to existing firms' prices and each new entrant undercuts the existing firm by a small but finite amount, price will eventually be driven down to the point at which profits are zero. Consideration of this frequently employed argument, which employs a price-conjectural variation different from the one used in Section II, is beyond the scope of this paper but our refutation of it can be found in our (1975b) paper.

unique equilibrium configuration. The argument would appear to run as follows: if firms are arranged in another configuration (say, rectangles), and just packed densely enough, i.e., a value of $h$ chosen, so that $R^a(h)=0$, they could all increase their profits by rearranging themselves into a hexagonal lattice. It is clear that if a rectangular configuration prevailed, all firms could increase their profits by agreeing collectively to rearrange themselves into a hexagonal configuration. It is equally clear from our models, however, that when each firm makes atomistic decisions based on conjectural variations with respect to other firms' locations and/or prices, there is no behavioral mechanism that will *necessarily* convert other configurations into hexagons (indeed we find it difficult to conceive of any behavioral mechanism suitable for atomistic decision taking that would be *sufficient* to do the job). Densest packing is not therefore an equilibrium condition of our models.

It seems to us, therefore, that many of the classic propositions of Löschian and post-Löschian location theory are suitable for analyzing optimal rules for collective decisions of a central-planning type, but are quite unsuitable for analyzing the outcome of decentralized decision making that is based on explicit conjectural-variation assumptions and that acts through the market mechanism.

## REFERENCES


M. Beckmann, *Location Theory,* New York 1968.

———, "Equilibrium Versus Optimum: Spacing of Firms and Patterns of Market Areas," *Northeast Reg. Sci. Rev.*, 1971, *1*, 1–20.

B. Bollabás and N. Stern, "The Optimal Structure of Market Areas," *J. Econ. Theory,* Apr. 1972, *4,* 174–79.

K. G. Denike and J. B. Parr, "Production in Space, Spatial Competition, and Restricted Entry," *J. Reg. Sci.*, Apr. 1970, *10,* 49–63.

B. C. Eaton, "Free Entry in One-Dimensional Market Models: Pure Profits and Multiple Equilibria," *J. Reg. Sci.*, forthcoming.

——— and R. G. Lipsey, (1972a) "The Principle of Minimum Differentiation Reconsidered," disc. pap. no. 87, Queen's Univ., 1972.

——— and ———, (1972b) "Unsuspected Perversities in the Theory of Location," disc. pap. no. 88, Queen's Univ., 1972.

——— and ———, (1975a) "The Principle of Minimum Differentiation Reconsidered: Some New Developments in the Theory of Spatial Competition," *Rev. Econ. Stud.*, Jan. 1975, *42,* 27–49.

——— and ———, (1975b) "Freedom of Entry and the Rate of Profit," study pap. no. 190, Queen's Univ.; study pap. no. 210, Univ. British Columbia 1975.

H. S. Grace, "Professor Samuelson on Free Enterprise and Economic Efficiency: A Comment," *Quart. J. Econ.*, May 1970, *84,* 337–40.

J. Hartwick, "Lösch's Theorem on Hexagonal Market Areas," *J. Reg. Sci.*, Aug. 1973, *13,* 213–21.

N. Kaldor, "The Irrelevance of Equilibrium Economics," *Econ. J.*, Dec. 1972, *28,* 1237–55.

A. Lösch, *The Economics of Location,* New Haven 1954.

E. Mills and M. Lav, "A Model of Market Areas with Free Entry," *J. Polit. Econ.*, June 1964, *72,* 278–88.

P. A. Samuelson, "The Monopolistic Competition Revolution," in R. E. Keunne, ed., *Monopolistic Competition Theory: Studies in Impact,* New York 1967.

# Input Choices and Uncertain Demand

By Duncan M. Holthausen*

The theory of the firm under demand uncertainty has to date assumed that the firm's input decisions are unaffected by the stochastic nature of demand.[1] This assumption is not usually made explicitly, but it is implied by the assumption that the firm's cost function, presumably derived by finding the minimum cost combination of inputs for any given level of output, is unaffected by the uncertainty. However, if firms must choose some of their inputs before the uncertainty regarding demand is resolved, then different input decisions may involve different levels of risk. Since firms' decisions regarding price and output are affected by demand uncertainty and non-linear risk preferences, as previous authors have shown, it seems reasonable to suspect that their input decisions will also be affected. But if input decisions are affected, then firms' cost functions are not invariant to uncertainty, and the results of previous studies must be reexamined.

Three different types of firms are considered in this paper: the competitive firm, the quantity-setting imperfect competitor, and the price-setting imperfect competitor. Input decisions of the competitive firm and the quantity-setting imperfect competitor are shown to be unaffected by demand uncertainty, and hence the assumption of an unchanging cost function is appropriate in these cases. For the price-setting imperfect competitor, however, input decisions are influenced by demand uncertainty, and an unchanging cost function is therefore not appropriate. This case is considered in some detail, and conditions under which results can differ from those previously reported in the literature are examined.

## I. Random Demand and Timing of Input Decisions

Following Leland, the implicit demand relationship

$$f(p, q, u) = 0 \tag{1}$$

is assumed to contain a random element $u$, which is not known *ex ante*, but has subjective probability density $dF(u)$. It is assumed that for any $u$, $p$ and $q$ are inversely related, and that larger values of $u$ are associated with greater demand.

Assuming $f$ to have continuous partial derivatives, (1) can be solved for either $p$ or $q$:

$$\begin{aligned} &p = p(q, u) \\ &\partial p(q, u)/\partial q < 0, \quad \partial p(q, u)/\partial u > 0 \end{aligned} \tag{2}$$

$$\begin{aligned} &q = q(p, u) \\ &\partial q(p, u)/\partial p < 0, \quad \partial q(p, u)/\partial u > 0 \end{aligned} \tag{3}$$

In both cases, the firm's revenue function is assumed to be concave.[2] For the perfectly competitive firm, relation (2) becomes simply

$$\begin{aligned} &p = p(u) \\ &\partial p(u)/\partial u > 0 \end{aligned} \tag{4}$$

and relation (3) does not exist.

The firm is assumed to produce its single output $q$ subject to the production function

* Assistant professor of business administration, The Amos Tuck School of Business Administration, Dartmouth College. This work was supported by a Tuck Associates research grant.

[1] See Hayne Leland, David Baron (1970, 1971), and Agnar Sandmo for recent work in this area.

[2] Specifically, $p(q,u)q$ is assumed concave in $q$, and $q(p,u)p$ is assumed concave in $p$.

$$(5) \qquad q = Q(K, L)$$

which gives output as a function of the two inputs, capital stock ($K$) and labor ($L$). The usual assumptions of positive marginal products $\partial Q/\partial L > 0$ and $\partial Q/\partial K > 0$ and diminishing returns $\partial^2 Q/\partial L^2 < 0$ and $\partial^2 Q/\partial K^2 < 0$ are also made. An important consideration is the firm's timing of input decisions. For the purposes of this paper, it is assumed that the level of capital stock must be decided before $u$ is known, and therefore before the actual level of demand is known. Thus capital stock is an *ex ante* control. Labor, on the other hand, is considered more flexible and may be adjusted after $u$ is known. Hence labor is an *ex post* control.

To reinforce this notion and to simplify the following analysis, the production function in (5) is solved for labor to give the labor-requirements function

$$(6) \qquad L = \mathcal{L}(q, K)$$

where $\mathcal{L}(q, K)$ gives the amount of labor necessary to produce output $q$ with capital stock $K$. This formulation emphasizes the assumption that labor is flexible and can be used to satisfy any level of demand. In fact, once the levels of $q$ and $K$ are determined, it is assumed that the firm's use of labor is determined by the labor-requirements function. Hence, labor is not considered a decision variable in the models that follow.

The signs of the first partial derivatives $\partial\mathcal{L}(q, K)/\partial q > 0$ and $\partial\mathcal{L}(q, K)/\partial K < 0$ follow from the assumption of positive marginal products for $K$ and $L$. The economic meaning of the first condition is that an increase in output (with capital stock held constant) can only be achieved by increasing labor, and the second condition implies that capital and labor can be substituted for each other to produce a given output.

In addition, $\partial^2\mathcal{L}(q, K)/\partial q^2 > 0$ follows from the assumption of diminishing returns to labor. In other words, with capital held constant, increasingly larger amounts of labor are required to produce each succeeding unit of output. It is also assumed that the isoquants of the production function are convex and therefore $\partial^2\mathcal{L}(q, K)/\partial K^2 > 0$. This condition is simply the usual one that as capital is substituted for labor in the production of a constant output, increasingly larger amounts of capital are required to replace each succeeding unit of labor. Or put another way, the marginal rate of technical substitution between capital and labor becomes larger (less negative) as more capital is used to produce a given output. Finally, the labor-requirements function is assumed to be convex in $q$ and $K$.[3]

## II. Optimal Input Choices for the Competitive Firm and the Quantity-Setting Imperfect Competitor

The firm's problem is taken to be the maximization of its expected utility of profit

$$(7) \qquad \underset{q,K}{\text{Maximize}}\ EU(\Pi)$$

where $U$ is a von Neumann-Morgenstern utility function with profit $\Pi$ its argument.

It is assumed that the firm purchases its inputs in perfectly competitive markets at constant prices, $w$ for labor and $c$ for capital. The firm's cost of capital (i.e., its opportunity cost for resources tied up in capital stock) is denoted $i$. Hence, profit is given by

$$(8) \quad \Pi = p(q, u)q - w\mathcal{L}(q, K) - icK$$

where $p(q, u) = p(u)$ for the perfect competitor.

The first-order conditions for (7) are

$$(9) \quad E\{[p(q, u) + q[\partial p(q, u)/\partial q] - w[\partial\mathcal{L}(q, K)/\partial q]]U'(\Pi)\} = 0$$

$$(10) \quad E\{[-w[\partial\mathcal{L}(q, K)/\partial K] - ic]U'(\Pi)\} = 0$$

[3] Note that although $\partial^2\mathcal{L}(q, K)/\partial q^2$ and $\partial^2\mathcal{L}(q, K)/\partial K^2$ are both positive, this is not enough to guarantee that $\mathcal{L}(q, K)$ is convex.

where $U'(\Pi)=dU(\Pi)/d\Pi>0$. Under the assumption that $p(q, u)q$ is concave and $\mathcal{L}(q, K)$ is convex, $\Pi$ is a concave function, and thus any point satisfying the first-order conditions provides a maximum for the problem as long as $U$ is linear or concave.[4]

It is important to notice that for the perfect competitor and the quantity-setting firm, both $q$ and $K$ are *ex ante* controls. That is, the firm sets the quantity to be produced and the amount of capital to be used before it knows the actual level of demand (i.e., before $u$ is known). Once $u$ is revealed, and assuming it sells all the $q$ it produces, the firm can charge any $p$ up to and including $\bar{p}=p(\bar{u}, \bar{q})$ where the bars denote the revealed value of $u$ and the setting of $q$. Being rational, the firm will set $\bar{p}$.

Since the firm sets both $q$ and $K$ *ex ante*, the amount of labor required, given by $L=\mathcal{L}(q, K)$, is known before $u$ is revealed. Thus $\mathcal{L}(q, K)$ is not stochastic, and (10) may be written as

$$(11) \quad [-w[\partial\mathcal{L}(q, K)/\partial K] - ic]EU'(\Pi) = 0$$

or

$$(12) \quad -\partial\mathcal{L}(q, K)/\partial K = \frac{ic}{w}$$

But (12) is just the deterministic rule for cost minimization. The firm employs capital and labor in such amounts that the marginal rate of substitution between inputs equals the ratio of their prices. Thus, the quantity-setting firm uses the least cost combination of capital and labor to produce the level of output it chooses. This should appear eminently reasonable, since the firm selects its level of output before the random element of demand $u$ is known. Thus $q$ is known for certain *ex ante*, and the firm may choose the level of $K$ (and hence $L$) which minimizes production cost.

It should be noted that although the quantity-setting firm facing stochastic demand uses the deterministic rule for cost minimization, it does not necessarily use the same levels of capital and labor as the firm operating under certainty does. As Leland, Baron (1970, 1971), and Sandmo have all shown, the perfect competitor or quantity setter facing uncertain demand will produce more or less than the firm under certainty depending on its risk preferences. Thus, the firm facing stochastic demand will generally use different amounts of capital and labor than an identical firm with certain demand.

Since the firm will always minimize costs for the level of output chosen, however, the assumption that the firm's cost function is unaffected by the presence of demand uncertainty is appropriate. Condition (9) may be analyzed to obtain results similar to those of Baron (1970), Leland, and Sandmo.

## III. Optimal Input Choices for the Price-Setting Imperfect Competitor

In the case of the price setter, the firm's problem is to

$$(13) \quad \underset{p,K}{\text{Maximize}}\ EU(\Pi)$$

where

$$(14) \quad \Pi = q(p, u)p - w\mathcal{L}(q(p, u), K) - icK$$

Writing $\mathcal{L}(q(p, u), K)$ as $\mathcal{L}(q, K)$ for notational convenience, the first-order conditions for (13) are

$$(15) \quad E\{[q(p, u) + p[\partial q(p, u)/\partial p] - w[\partial\mathcal{L}(q, K)/\partial q] \cdot [\partial q(p, u)/\partial p]]U'(\Pi)\} = 0$$

$$(16) \quad E\{[-w[\partial\mathcal{L}(q, K)/\partial K] - ic]U'(\Pi)\} = 0$$

In this case, $p$ and $K$ are the firm's *ex ante* controls. The firm chooses its price and amount of capital before the level of

[4] The revenue function $p(u)q$ for the competitive firm, is linear, but $\Pi$ is still concave because of the production function assumptions.

demand is known. Once the uncertain element $u$ of demand is realized, the firm produces the $q$ implied by $\bar{q}=q(\bar{u}, \bar{p})$, where the bars represent the realization of $u$ and the setting of $p$. Although the firm might decide to produce and sell less than $\bar{q}$, this would be rational only if the marginal cost of production at $\bar{q}$ exceeded the price $\bar{p}$. To rule out this possibility, it is assumed that price is greater than the marginal cost of production, i.e., $\bar{p}>w[\partial\mathcal{L}(\bar{q}, \bar{K})/\partial q]$, for the chosen values of $p$ and $K$ and the realized value of $u$.

Using the fact that for two random variables $X$ and $Y$, $E(XY)=E(X)E(Y)+cov\,(X, Y)$, equation (16) may be written as

$$(17)\quad E\{-w[\partial\mathcal{L}(q, K)/\partial K] - ic\}EU'(\Pi) + w\,cov\,[-\partial\mathcal{L}(q, K)/\partial K, U'(\Pi)] = 0$$

or

$$(18)\quad E[-\partial\mathcal{L}(q, K)/\partial K] = \frac{ic}{w} - \frac{cov\,[-\partial\mathcal{L}(q, K)/\partial K, U'(\Pi)]}{EU'(\Pi)}$$

In this case, the firm departs from the expected cost minimizing strategy as long as the covariance term is nonzero. A sufficient condition for the covariance term to be zero is that the firm's utility be linear in profits, i.e., that the firm be risk neutral. In this case, $U'(\Pi)$ is constant and the covariance must be zero. The risk-neutral firm, therefore, sets its expected marginal rate of technical substitution equal to the slope of its isocost line and thereby minimizes expected cost.

Firms with non-linear risk preferences will depart from this expected cost minimizing strategy and hence will be in effect operating with a different cost function than the risk neutral firm. To determine the direction of the effect of risk preferences, the sign of the covariance term must be determined. It is shown in the next section that this depends on the normality of capital and on the profitability of increased output with price held fixed.

## IV. Risk Preferences and Departures from Minimum Cost for Price-Setting Firms

The sign of the covariance term in (18) can be determined by examining the changes in $-\partial\mathcal{L}(q, K)/\partial K$ and $U'(\Pi)$ as $u$ changes. This requires that we be able to sign the terms

$$(19)\quad \partial U'(\Pi)/\partial u = [\partial\Pi/\partial u]U''(\Pi) = [p - w[\partial\mathcal{L}(q, K)/\partial q]][\partial q(p, u)/\partial u]U''(\Pi)$$

and

$$(20)\quad \partial[-\partial\mathcal{L}(q, K)/\partial K]/\partial u = -[\partial^2\mathcal{L}(q, K)/\partial q\partial K][\partial q(p, u)/\partial u]$$

As shown in the Appendix, under the assumption that price exceeds the marginal cost of producing additional units, $\partial\Pi/\partial u$ is positive and hence the sign of (19) is the same as the sign of $U''(\Pi)$. Also, if capital is a normal input (which would usually be the case), then the sign of (20) is positive. Hence under these assumptions it follows that

$$(21)\quad cov\,[-\partial\mathcal{L}(q, K)/\partial K, U'(\Pi)] \begin{cases} >0 & \text{for } U \text{ convex} \\ =0 & \text{for } U \text{ linear} \\ <0 & \text{for } U \text{ concave} \end{cases}$$

Applying (21) to equation (18) yields

$$(22)\quad E[-\partial\mathcal{L}(q, K)/\partial K] \begin{cases} <\dfrac{ic}{w} & \text{for } U \text{ convex} \\ =\dfrac{ic}{w} & \text{for } U \text{ linear} \\ >\dfrac{ic}{w} & \text{for } U \text{ concave} \end{cases}$$

It follows that risk-preferring firms ($U$ convex) use a greater amount of capital than that which minimizes expected cost for the level of output produced while risk-averse ($U$ concave) firms use a smaller amount of capital than the expected cost

minimizing amount for the level of output produced.[5]

In terms of a cost function such as $c(q)+F$ where $c(q)$ represents variable cost and $F$ fixed cost, both $c(q)$ and $F$ will be affected. The risk-preferring firm will use a higher level of fixed cost than the risk-averse firm, since its capital stock is large relative to its output, while the risk-averse firm's capital stock is small relative to output. On the other hand, the risk-averse firm will require relatively more labor than the risk-preferring firm, and so its variable costs will be higher. Thus, the assumption that the firm's cost function is unaffected by demand uncertainty and risk preferences is incorrect.

The finding that the risk-averse firm uses a production process with low fixed costs and high variable costs while the risk-preferring firm uses a high fixed cost, low variable cost process is intuitively appealing. The low fixed cost, high variable cost process used by the risk-averse firm gives it a low break-even point, and keeps its losses small should demand turn out to be less than expected. Doing this, of course, does reduce the profit that the firm can make if demand is high, because it will have to use an inefficiently low capital-labor ratio in that case.[6] These are the kinds of tradeoffs typical of risk-averse behavior; the firm gives up the possibility of large profits to insure against large losses.

The risk-preferring firm, on the other hand, uses a higher degree of leverage than either the risk-neutral or risk-averse firm. By using a high fixed cost production process, it incurs large losses if demand is low, but reaps large profits if demand is greater than expected.

## V. The Effect of Increasing Risk Aversion on the Optimal Level of Capital Stock for Price-Setting Firms

The results of Section IV can be generalized for the case of the risk-averse firm by showing that for a given price the firm will reduce its optimal level of capital stock as aversion to risk increases. The degree of risk aversion will be measured by the Pratt-Arrow index, which is defined as

$$r(\Pi) = -\frac{U''(\Pi)}{U'(\Pi)} \tag{23}$$

This measure of risk aversion is always positive for a risk-averse firm since $U''(\Pi)<0$. Defining the concept of "increasing risk aversion" to mean that the firm would be willing to pay an increasing amount to insure against a given risk, John Pratt has shown that the index $r(\Pi)$ increases as the utility function becomes more risk averse.

Using this measure of risk aversion, it is now shown that with price held fixed, the more risk averse a firm becomes, the less capital stock it will use. Consider two firms with concave utility functions $U_1(\Pi)$ and $U_2(\Pi)$ with $r_1(\Pi)>r_2(\Pi)>0$ for all $\Pi$, where $r_1$ and $r_2$ are the Pratt-Arrow indexes of risk aversion for $U_1$ and $U_2$, respectively. If price is fixed at the same level for both firms, firm one's first-order condition is given by

$$\int_{-\infty}^{\infty}\{[-w[\partial\mathcal{L}(q, K_1^*)/\partial K] - ic]U_1'(\Pi)\}\,dF(u) = 0 \tag{24}$$

where $dF(u)$ is the subjective probability density of the random element of demand,

[5] Again, it should be noted that the level of output produced by firms with non-linear risk preferences will usually differ from that produced by risk-neutral firms or firms operating in a deterministic world. This topic is treated in Sections VI and VII.

[6] It may be noted in passing that since risk-averse price-setting firms have been shown to undercapitalize on average, and since regulated firms are characterized by price-setting behavior, the Averch-Johnson overcapitalization bias that accompanies rate-of-return regulation may be beneficial if regulated firms are risk averse. See my unpublished paper for further elaboration of this point.

and $K_1^*$ is firm one's optimal level of capital stock. As shown in the Appendix, $-\partial \mathcal{L}(q, K)/\partial K$ is increasing in $u$ if capital is a normal input. Thus, $-w[\partial \mathcal{L}(q, K_1^*)/\partial K]-ic$ could be negative for small $u$ and positive for large $u$. Define $u^0$ as the value of $u$ for which $-w[\partial \mathcal{L}(q, K_1^*)/\partial K]-ic=0$. Then $-w[\partial \mathcal{L}(q, K_1^*)/\partial K]-ic$ will be negative for $u<u^0$ and positive for $u>u^0$. Also define $\Pi^0$ as the value of $\Pi$ when $u=u^0$.

Partitioning (24) into two parts and dividing by the constant $U_1'(\Pi^0)$, the firm's first-order condition may be written as

$$(25)\quad \int_{-\infty}^{u^0} \frac{U_1'(\Pi)}{U_1'(\Pi^0)}[-w[\partial \mathcal{L}(q, K_1^*)/\partial K] - ic]dF(u) + \int_{u^0}^{\infty} \frac{U_1'(\Pi)}{U_1'(\Pi^0)}[-w[\partial \mathcal{L}(q, K_1^*)/\partial K] - ic]dF(u) = 0$$

Now evaluate firm two's first-order condition at the optimal point for firm one (i.e., at $K_1^*$) and divide this condition by the constant $U_2'(\Pi^0)$ to obtain

$$(26)\quad \int_{-\infty}^{u^0} \frac{U_2'(\Pi)}{U_2'(\Pi^0)}[-w[\partial \mathcal{L}(q, K_1^*)/\partial K] - ic]dF(u) + \int_{u^0}^{\infty} \frac{U_2'(\Pi)}{U_2'(\Pi^0)}[-w[\partial \mathcal{L}(q, K_1^*)/\partial K] - ic]dF(u)$$

Subtracting (25) from (26) yields

$$(27)\quad \int_{-\infty}^{u^0}\left[\frac{U_2'(\Pi)}{U_2'(\Pi^0)} - \frac{U_1'(\Pi)}{U_1'(\Pi^0)}\right]\cdot[-w[\partial \mathcal{L}(q, K_1^*)/\partial K] - ic]dF(u) + \int_{u^0}^{\infty}\left[\frac{U_2'(\Pi)}{U_2'(\Pi^0)} - \frac{U_1'(\Pi)}{U_1'(\Pi^0)}\right]\cdot[-w[\partial \mathcal{L}(q, K_1^*)/\partial K] - ic]dF(u)$$

One of Pratt's results is required at this point. Pratt's Theorem 1 shows, among other things, that if $U_1(\Pi)$ is more risk averse than $U_2(\Pi)$ in the sense that $r_1(\Pi)\geq r_2(\Pi)$ for all $\Pi$ [and $>$ for at least one $\Pi$], then[7]

$$\frac{U_1'(x)}{U_1'(w)} \leq [<] \frac{U_2'(x)}{U_2'(w)} \qquad \text{for } w < x$$

As shown in the Appendix, $\Pi$ is increasing in $u$ as long as price exceeds marginal cost, and thus $\Pi\leq\Pi^0$ for $u\epsilon(-\infty, u^0)$ and $\Pi\geq\Pi^0$ for $u\epsilon(u^0, \infty)$. Therefore, the first term in brackets in (27) is negative by Pratt's equation 20, and the second term in brackets is negative by definition of $u^0$. Hence, the first integral in (27) is positive. The second integral in (27) is also positive, since the first term in brackets is positive by Pratt's equation 20, and the second term is positive by definition of $u^0$. Hence, (27) is positive, which implies (26) is also positive. But (26) is the first-order condition for firm two evaluated at $K_1^*$. Under the concavity assumptions made earlier, $K$ must be increased above $K_1^*$ to drive (26) to zero. Thus, it follows that $K_2^*>K_1^*$. The more risk-averse firm (firm one) uses less capital.

Hence, with price held fixed, the amount of capital stock used decreases as risk aversion increases. Thus, the more risk-averse firm will use more labor to produce any given level of output and hence will use a lower capital-labor ratio. Since a risk-neutral firm uses the efficient ratio, any risk-averse firm will undercapitalize as noted before.

Under the hypothesis of decreasing absolute risk aversion, it follows that the risk-averse firm's capital-labor ratio increases as its wealth increases. Loosely interpreted, this implies that the larger the risk-averse firm becomes, the closer to the efficient capital-labor ratio it will operate. Hence, one might expect larger firms to

[7] See Pratt's equation 20, p. 129.

operate more efficiently than smaller ones, *ceteris paribus*.

### VI. The Effect of Increasing Risk Aversion on the Optimal Price for Price-Setting Firms

Results similar to those of Leland can be obtained from this model if it is assumed that capital stock is held constant as the level of risk aversion varies. Holding $K$ fixed at, say, $\overline{K}$ is equivalent to assuming the cost function is invariant with changes in risk aversion, since the total cost function will be $w\mathcal{L}(q, \overline{K})+ic\overline{K}$ which is of the form $c(q)+F$. Using a method of analysis similar to that in Section V, it can be shown that if $K$ is fixed, then as aversion to risk increases, price decreases (increases) (is constant) as $\partial(\partial\Pi/\partial p)/\partial u>0$ $(<0)(=0)$. It follows, therefore, that the risk-averse firm sets its price below the risk-neutral firm if $\partial(\partial\Pi/\partial p)/\partial u>0$, above if $\partial(\partial\Pi/\partial p)/\partial u<0$, and equal to the risk-neutral firm if $\partial(\partial\Pi/\partial p)/\partial u=0$.

The sign of $\partial(\partial\Pi/\partial p)/\partial u$ cannot be determined unambiguously a priori. For simple demand functions, its sign is usually positive. For example, it is positive if $u$ appears only as an additive shift variable. However, in general its sign is ambiguous. Thus the effect of uncertainty on the price charged by the risk-averse price setter (with capital stock fixed) is ambiguous. It depends critically on how changes in uncertainty affect the profitability of price changes as Leland has shown.

### VII. Simultaneous Effects on Price and Capital Stock

Examining the effects of uncertainty on capital and price separately may give some insight into the problem, but a complete analysis requires simultaneous treatment of capital and price. As is often the case in economics, such treatment is not entirely satisfactory, but it does indicate that the earlier results on price given in Section VI are not complete as they stand. Allowing the firm to make both input and price decisions simultaneously changes its pricing decision in some cases.

It is necessary to introduce a bit more notation at this point. Let $\phi\equiv EU(\Pi)$. Then $\phi_p=\partial EU(\Pi)/\partial p$, $\phi_{pK}=\partial^2 EU(\Pi)/\partial p\partial K$, and so on. Also, let $r$ be the Pratt-Arrow index of risk aversion, so that, for example, $dp/dr$ measures the change in price that occurs with a change in risk aversion.

We are interested in the changes in price and capital that accompany changes in risk aversion, so the signs of $dp/dr$ and $dK/dr$ are of interest. The relevant comparative statics experiment may be expressed as

$$\text{(28)}\qquad \begin{bmatrix}\phi_{pp} & \phi_{pK}\\ \phi_{pK} & \phi_{KK}\end{bmatrix}\begin{bmatrix}dp/dr\\ dK/dr\end{bmatrix} = -\begin{bmatrix}\phi_{pr}\\ \phi_{Kr}\end{bmatrix}$$

In Section V it was shown that under certain reasonable conditions when price is fixed, the level of capital stock declines as risk aversion increases. Hence $dK/dr\,|_{p\text{ constant}}<0$. Taking the total differential of the first-order condition $\phi_K=0$, we have (for $p$ fixed)

$$\text{(29)}\qquad \phi_{KK}dK + \phi_{Kr}dr = 0$$

Therefore,

$$\text{(30)}\qquad dK/dr\,|_{p\text{ constant}} = -\frac{\phi_{Kr}}{\phi_{KK}}$$

Since $dK/dr\,|_{p\text{ constant}}<0$ and $\phi_{KK}<0$ by second-order conditions, it follows that $\phi_{Kr}<0$. Likewise, it can easily be shown that $\phi_{pr}\lesseqgtr 0$ as $\partial(\partial\Pi/\partial p)/\partial u\gtreqless 0$.

The only other sign required to carry out the comparative statics experiment is that of $\phi_{pK}\equiv\partial^2 EU(\Pi)/\partial p\partial K$, which may be written

$$\begin{aligned}\text{(31)}\qquad \phi_{pK} = E\{&-w[\partial^2\mathcal{L}(q, K)/\partial q\partial K]\\ &\cdot[\partial q(p, u)/\partial p]U'(\Pi)\\ &+[-w[\partial\mathcal{L}(q, K)/\partial K] - ic]\\ &\cdot[\partial\Pi/\partial p]U''(\Pi)\}\end{aligned}$$

As shown in the Appendix, $\partial^2 \mathcal{L}(q, K)/\partial q \partial K < 0$ as long as capital is a normal input. Since $\partial q(p, u)/\partial p < 0$ and $U'(\Pi) > 0$ by assumption, the first term in (31) is negative. If the firm is risk neutral, $U''(\Pi) = 0$ and $\phi_{pK}$ is always negative. For a risk-averse firm $U''(\Pi) < 0$, but the signs of $[-w[\partial \mathcal{L}(q, K)/\partial K] - ic]$ and $\partial \Pi/\partial p$ are ambiguous. It would seem reasonable to assume, however, that for small degrees of risk aversion ($U''(\Pi)$ close to zero), $\phi_{pK}$ would still be negative, although as risk aversion increases we cannot rule out the possibility that $\phi_{pK}$ becomes positive. The assumption $\phi_{pK} < 0$ will be used in what follows, but its tenuous nature should be remembered.

Now solve (28) for $dp/dr$ and $dK/dr$ as follows.

$$(32) \qquad dp/dr = \frac{-\phi_{pr} - \phi_{pK}(dK/dr)}{\phi_{pp}}$$

$$(33) \qquad dK/dr = \frac{\phi_{pK}\phi_{pr} - \phi_{Kr}\phi_{pp}}{\phi_{pp}\phi_{KK} - \phi_{pK}^2}$$

The denominator of (33) is positive by second-order conditions. The second term in the numerator is negative since $\phi_{Kr} < 0$ as shown above and $\phi_{pp} < 0$ by second-order conditions. If $\partial(\partial \Pi/\partial p)/\partial u \leq 0$, then $\phi_{pr} \geq 0$ as shown above, and under the assumption $\phi_{pK} < 0$, it follows that $dK/dr < 0$. Using this in (32), it follows that $dp/dr > 0$. If $\partial(\partial \Pi/\partial p)/\partial u > 0$, then $\phi_{pr} < 0$ and the sign of $dK/dr$ is ambiguous. In this case, we cannot have both $dK/dr > 0$ and $dp/dr > 0$, but any other pair of signs is acceptable.

These results are generally consistent with the nonsimultaneous results reported earlier in this paper. For example, if $\partial(\partial \Pi/\partial p)/\partial u < 0$, then $dK/dr < 0$ and $dp/dr < 0$. The risk-averse firm uses less capital and sets a higher price as risk aversion increases. In particular, it uses less capital and a higher price than a risk-neutral firm. These results are identical to those presented in Sections V and VI.

If the profitability of a price increase is unaffected by uncertainty (i.e., $\partial(\partial \Pi/\partial p)/\partial u = 0$), the simultaneous result is $dK/dr < 0$ and $dp/dr > 0$. However, in this case, the result differs from the earlier finding that $dp/dr = 0$. In this instance, then, considering pricing and input decisions simultaneously is important. The risk-averse firm will use less capital than the risk-neutral firm and will set a higher price. Leland's model and the analysis presented in Section VI of this paper both assume that the firm's input decisions are not affected by uncertainty, and both lead to the conclusion that if $\partial \Pi/\partial p$ is independent of $u$, the risk-averse firm will set the same price as the risk neutral firm. The simultaneous result shows, however, that the risk-averse firm will adjust its capital stock downward and hence raise its price above the risk neutral price.

Finally, for $\partial(\partial \Pi/\partial p)/\partial u > 0$, the results are ambiguous. We can rule out the possibility that the risk-averse firm will simultaneously increase its capital stock and price above the risk neutral level. Other types of behavior are, however, consistent with the model. For example, the firm could decrease capital stock and increase price as it became more risk averse. Thus, while the nonsimultaneous result implies $dp/dr < 0$, a simultaneous analysis shows that $dp/dr > 0$ cannot be ruled out when capital stock is decreasing in risk aversion.

## VIII. Conclusion

A model of an expected utility maximizing firm facing random demand has been developed. The major difference between this model and earlier models reported in the literature is that the firm is allowed to make input decisions as well as pricing and output decisions.

It was shown that the results of earlier models are correct for the perfectly competitive firm and the quantity-setting im-

perfect competitor. In both of these cases, the firm sets the quantity it is to produce before the random element of demand is known. Hence, quantity is an *ex ante* control, and the firm selects the optimal input combination as it would if there were no uncertainty.

For the price-setting imperfect competitor, however, quantity is an *ex post* control. That is, the quantity to be produced is uncertain at the time some input decisions must be made. Here, the firm's attitude toward risk plays an important role in its behavior. In general, the risk-averse or risk-preferring firm will use a different input mix than the risk-neutral firm. Hence, earlier models which have assumed the firm's cost function to be invariant with changes in risk preference are not complete.

The major findings for the price-setting imperfect competitor are:

1. The risk-averse firm will use an expected capital-labor ratio less than the efficient ratio while a risk-preferring firm will use an expected capital-labor ratio greater than the efficient ratio.
2. If price is fixed, the amount of capital stock used by the risk-averse firm decreases as aversion to risk increases.
3. If capital stock is held fixed, the price charged by the risk-averse firm may decrease, increase, or remain constant depending on how the profitability of a price increase is affected by uncertainty.
4. Considering both price and capital stock decisions simultaneously, some of the earlier results are shown to be incorrect or not complete. The results affected are primarily those dealing with pricing behavior.

In general, then, earlier models of the price-setting firm under demand uncertainty are incomplete because they assume the firm's cost function to be a given. When both prices and inputs are considered simultaneously, some deviations from the results of the earlier models occur, and in some cases, results that had been unambiguous are now not as clear cut.

## Appendix

The sign of (19) depends on the sign of $p-w[\partial\mathfrak{L}(q,K)/\partial q]$, since $\partial q(p,u)/\partial u>0$ and the sign of $U''(\Pi)$ depends on the concavity-convexity of $U$. The term $\partial\mathfrak{L}(q,K)/\partial q$ represents the additional labor required to produce a small increment in output. Hence, $w[\partial\mathfrak{L}(q, K)/\partial q]$ is the marginal cost of producing an increment in output when the amount of capital is fixed. Under the assumption made in Section III, that price is greater than the marginal cost of production (a reasonable condition in most cases), $p-w[\partial\mathfrak{L}(q,K)/\partial q]$ is positive, and the sign of (19) is the same as the sign of $U''(\Pi)$.

The sign of (20) is the same as the sign of $\partial^2\mathfrak{L}(q,K)/\partial q\partial K$, since $\partial q(p,u)/\partial u$ is positive. The sign of $\partial^2\mathfrak{L}(q,K)/\partial q\partial K$ in turn depends on whether capital is a normal or an inferior input. The assumption that capital is a normal input appears to be an entirely reasonable one for most firms and will be used in what follows. Under this assumption, larger amounts of capital would be used as planned output increases.

To sign $\partial^2\mathfrak{L}(q,K)/\partial q\partial K$, we must ask how the marginal rate of substitution between capital and labor $(\partial\mathfrak{L}(q,K)/\partial K)$ behaves in the short run as output increases and capital is held constant. We know, since capital is assumed to be a normal input and input prices are assumed constant, that in the long run capital would be increased as output increases and that $\partial\mathfrak{L}(q,K)/\partial K$ would remain constant, because any optimal point must be a point of tangency with an isocost line whose slope is constant. Therefore, if output is increased and capital is held fixed, the firm will have to operate with less than the desired amount of capital. We know that $\partial\mathfrak{L}(q, K)/\partial K$ will decrease (i.e., become more negative) when capital is reduced because of the assumption of convex isoquants (which im-

plies $\partial^2\mathcal{L}(q, K)/\partial K^2>0$). Hence $\partial\mathcal{L}(q, K)/\partial K$ will be smaller than it would have been had the firm been able to adjust its level of capital stock to the increased level of output. It follows that $\partial^2\mathcal{L}(q, K)/\partial q\partial K<0$, and therefore (20) is positive under the assumption that capital is a normal input.

## REFERENCES


K. J. Arrow, *Essays in the Theory of Risk Bearing,* Chicago 1971.

H. Averch and L. L. Johnson, "Behavior of the Firm Under Regulatory Constraint," *Amer. Econ. Rev.,* Dec. 1962, *52,* 1052–69.

D. Baron, "Price Uncertainty, Utility, and Industry Equilibrium in Pure Competition," *Int. Econ. Rev.,* Oct. 1970, *11,* 463–80.

———, "Demand Uncertainty in Imperfect Competition," *Int. Econ. Rev.,* June 1971, *12,* 196–208.

D. M. Holthausen, "Uncertainty, Risk Preferences and the Behavior of the Regulated Firm," unpublished paper, 1974.

H. E. Leland, "Theory of the Firm Facing Uncertain Demand," *Amer. Econ. Rev.,* June 1972, *62,* 278–91.

J. W. Pratt, "Risk Aversion in the Small and in the Large," *Econometrica,* Jan.-Apr. 1964, *32,* 122–36.

A. Sandmo, "On the Theory of the Competitive Firm Under Price Uncertainty," *Amer. Econ. Rev.,* Mar. 1971, *61,* 65–73.

# The Long-Run Analysis of the Labor-Managed Firm: An Alternative Interpretation

*By* EIRIK G. FURUBOTN*

The pure theory of the labor-managed firm, as developed by Benjamin Ward, Evsey Domar, and Jaroslav Vanek,[1] is based on the assumption that the collective will seek to maximize income per worker. Apart from the emphasis on this special behavioral postulate, however, the analysis runs parallel with conventional micro theory.[2] Clearly, the approach has some advantages. In so far as the firm's objective is taken to be simple "wage" maximization, a straightforward optimization problem can be formulated; and, then, by following familiar technical procedures, it is possible to determine the equilibrium levels of factor usage in the long or short run, the response of the firm to price changes, etc.[3] The question that arises, though, is whether this particular interpretation of the case leads to any significant understanding of the behavior of the labor-managed enterprise. The Ward-Domar-Vanek model may have surface plausibility, but the model's crucial element, the maximand, can certainly be challenged.[4]

Reduced to simplest terms, the argument to be developed here is that an adequate theory of the self-managed firm must consider the preferences and wealth increasing opportunities of those individuals *actually making economic decisions*. At any moment of time, particular individuals are members of the working collective, and it is this constituency that must decide the firm's actions in the next time frame. The precise set of policy alternatives open depends on various factors, including the existing structure of property rights; but whatever the alternatives present, policies will presumably be evaluated in light of their probable consequences for the specific group involved in making the decisions. The worker-voters who actually determine the firm's behavior are not concerned with the position of labor in general, but with their own welfare.

More than semantics is at issue. For realistically a policy that maximizes income per laborer (without regard to the particular workers involved) need not be optimal from the standpoint of the "original" members of the collective. This situation carries direct implications for the theory of the firm; the behavior predicted for the labor-managed enterprise will take one form when the maximand is income per laborer, and an entirely different form when the maximand is the welfare index of a particular group of workers. Existing literature treats the first case. But the second case, which seems much more in harmony with the concept of untrammeled labor management, requires further study and is the concern of the present investigation.

* Professor of economics, Texas A&M University. The writing of this paper was facilitated by a grant from the National Science Foundation.

[1] See also Abram Bergson, A. B. Atkinson, the author (1971, 1974), and S. Charles Maurice and Charles Ferguson.

[2] See Vanek (1969, pp. 1006–07).

[3] See Vanek (1970, ch. 2, pp. 39–41) and Maurice and Ferguson, pp. 21–27.

[4] There is no doubt that the choice of the maximand is crucial to the model of the labor-managed firm. See Maurice and Ferguson, pp. 18–19.

## I. Labor-Management and the Collective's Preference Function

The philosophy of labor-management presupposes self-government by the workers and assumes that each firm will be guided by the desires of its employees.[5] This emphasis on freedom from external controls is highly significant; indeed, the labor-managed firm is of special interest to socialist nations precisely because the new organization promises to end the alienation of labor. It follows, of course, that if alienation is to be countered, the firm's policies must always be arrived at through democratic processes in which all current employees take part. The working collective is best understood as a microcosmic society, and the political choices that are made within this society are crucial. That is, the voting process decides control of the firm and ultimately leads to the establishment of the firm's objectives with respect to wages, employment, output, investment, etc. Like other societies, the collective is an ongoing social system and subject to change. Both the characteristics of the collective and the external conditions faced by it will tend to shift over time. Thus, in formulating an optimization model that considers how the labor-managed firm adjusts to economic forces, it is essential to use dynamic analysis and conceive of the equilibration process as an ordered sequence of events.

As a first step, attention can be given to the technological constraints under which the firm operates through time. For convenience, we assume that the firm produces subject to a simple stock-flow function:

$$\begin{aligned} &(1) \qquad & q_t &= f(L_t, K_t) \\ && R_t &= g(K_t) \end{aligned}$$

Here $q_t$ represents the flow of commodity output in any period t, $L_t$ is the contemporaneous flow of labor services, and $K_t$ is the stock of capital equipment actually available for production in period t. The time intervals considered are finite and, by assumption, the physical capital acquired in one period cannot be used in active production until the next period. Finally, the raw material input (or inputs) $R_t$ is indicated in (1) as a factor shadow; thus, the amount used per period is always related to the magnitude of the capital stock in accordance with the stipulations of function $g$.

Given system (1), the firm's initial conditions can be stated as the magnitudes of the labor and capital inputs available to the collective at the first time period studied (t=1). These data ($L_1$, $K_1$) are important because *inter alia* they establish: (i) the specific group of workers that is eligible to vote on the firm's policies for the next period,[6] and (ii) the basic economic position from which the group must commence productive activity. All current members of the collective ($L_1$) are free to vote, but individual differences among workers in respect to such things as risk preference, time preference, length of planning horizon, desire for mobility, etc., must lead to different views on what actions the firm should take. Which view prevails depends on the outcome of the political process that is carried on within the firm. Assuming, however, that a relatively homogeneous majority can be formed in period one, it is this group that determines the policies of the firm for period two.

The workers controlling the firm at any time have incentive to perpetuate their domination of policy and, in such way, preserve the best possible economic and social conditions for themselves. Presum-

[5] Branko Horvat is emphatic about this point, p. 99, pp. 106–07.

[6] Assuming that each worker provides the same (fixed) number of hours of labor service per period, any flow $L_t$ implies a unique corresponding employment level for the firm.

ably, then, the *original majority* (as of t=1) will attempt to retain power; and if this group can coalesce into a permanent interest bloc and sustain itself politically over time, the basis of decision-making behavior by the collective is established. In effect, the firm will be guided by a criterion function reflecting the preferences of the employees forming the successful "original majority." Thus, for purposes of the model, it is possible to write the multiperiod utility index:

$$(2)\quad U = \phi(C_1, C_2, \ldots, C_T, X_{11}, X_{12}, \ldots, X_{1T}, X_{21}, X_{22}, \ldots, X_{2T}, \ldots, X_{s1}, X_{s2}, \ldots, X_{sT})$$

Equation (2) is a type of community welfare function and its derivation poses certain theoretical difficulties. It can be assumed, however, that the majority workers have broadly similar tastes so that there is agreement on: (i) the length of the planning horizon (T); (ii) the general time shape of the disposable income stream that is to be generated for the members of the collective ($C_1, C_2, \ldots, C_T$); and (iii) the economic and social environment that is to be maintained within the firm during the period its planned operations are being carried out ($X_{11}, X_{12}, \ldots, X_{21}, \ldots, X_{s1}, \ldots, X_{sT}$). Given agreement on issues (i)–(iii), a basis for rational policy making does exist.[7] On the practical level, at least, preference aggregation is possible,[8] and a consistent long-run program representing the interests of the original majority can be formulated.

There is no problem in understanding that decision makers will be concerned with the income stream the firm can provide ($C_t$, t=1, 2, ..., T). But, under labor-management, workers will also have interest in and control over the internal environment of the firm at each period to the horizon. Many factors, of course, influence the firm's working atmosphere or environment—including such things as the strictness of labor discipline, the pace of work, the levels of risk assumed, the absolute size of the collective, etc. In (2), forces of this type are indicated by the variables: $X_{kt}$, $k=1, 2, \ldots, s$, t=1, 2, ..., T. It is suggested that there are $s$ distinct dimensions to the firm's environment that can be measured with appropriate indexes. In general, then, the satisfaction level of the original majority $U$ is determined by the magnitudes of the variables $C_t$, $X_{kt}$; and each proposed policy action of the firm is judged as relatively desirable or less desirable, depending on how it influences these variables and the value of the index $U$.

Since the size of the firm can change over time, it is entirely possible that the labor force will expand or contract. From the standpoint of the original majority, however, the decision on an ideal employment program presents some difficulty. Any movement from the initial position $L_1$ must be considered carefully because change in the size of the collective yields simultaneous effects in three quite different areas. For each change, the original majority has to assess the costs and benefits associated with: (i) a *production* effect, (ii) an *environmental* effect, and (iii) a *political* effect. For example, in the case of expansion, the decision makers must estimate not only the physical productivity of any additional labor hired, but the impact the labor increment will have on the firm's working environment and on the firm's internal political structure. The last consideration is particularly important since change in the membership of the collective normally implies some danger to the interests of the original majority. New workers coming into a labor-managed enterprise are not just factors of production but po-

[7] See Roy Radner, pp. 47–53.

[8] A majority voting process can be expected to lead to a determinate solution for the production-investment-employment plan. See Gordon Tullock, p. 257.

tential policy makers as well. Thus, the way is open for a transfer of power; voting patterns can be shifted because of the added workers, and new policy directions established for the firm that are contrary to the desires of the original majority.

As far as the environmental variables are concerned, it is clear that these quality indexes ($X_{kt}$) are necessarily tied to the employment level. Each period, the social-economic conditions to be maintained within the firm are decided by a majority vote of the existing members of the collective. In planning the ideal labor program ($L_2, L_3, \ldots, L_T$), then, the *original majority* must anticipate both the direct effects of "crowding" on the firm's environment and the political consequences of labor change. Relative to the latter, what is needed is a series of estimates showing how a labor force of any given size will vote on the issues represented by the $X_{kt}$ variables. The general situation of the decision makers is easily summarized. Assuming that all of the direct and indirect effects of labor variation can be assessed, any proposed value of $L_t$ ($t=2, 3, \ldots, T$) implies a corresponding value for each environmental variable. Specifically, we can write:

$$X_{kt} = \theta_{kt}(L_t), \qquad L_t \geq L_1 \tag{3}$$
$$k = 1, 2, \ldots, s \quad t = 2, 3, \ldots, T$$

These conversion relations are obviously subjective in nature; they refer to what the initial majority *thinks* will happen as the labor force is expanded to any level in any period. Precisely how the respective translations work out in quantitative terms depends on the psychological attitudes of the decision makers, on the particular estimation procedures employed, and on the initial conditions prevailing in the collective—including the size of the collective in period one and the size of the original majority group. Various outcomes are conceivable, but it seems reasonable to conjecture that the members of the original majority will be relatively cautious and suspicious of change. Any enlargement of the collective tends to make their political control less secure and promises some deterioration of the firm's environment. Thus, an *inverse* relation can be expected to exist between each environmental index $X_{kt}$ and the contemporaneous labor input ($L_t$).

From a formal standpoint, it will be convenient to substitute (3) in (2) and secure a new objective function written in terms of $C_t$ and $L_t$.

$$U = \phi[C_t, \theta_{kt}(L_t)], \quad t = 1, 2, \ldots, T \tag{4}$$
$$k = 1, 2, \ldots, s$$

This version of the utility index makes it easy to understand why there is a general reluctance on the part of the original majority to increase the size of the labor force. By differentiation, we know:

$$\frac{\partial U}{\partial L_t} = \frac{\partial U}{\partial C_t}\frac{\partial C_t}{\partial L_t} + \sum_{k=1}^{s} \frac{\partial U}{\partial X_{kt}}\frac{\partial X_{kt}}{\partial L_t}$$

The magnitude of $C_t$ at any period depends directly on the physical productivity of labor; and under technical conditions where the firm's *earnings per worker* can be increased through an enlargement of the labor force,[9] the term $\partial U/\partial C_t \, \partial C_t/\partial L_t$ will be positive. But according to (3), the other terms deciding the value of $\partial U/\partial L_t$ are always negative. Thus, the positive productivity effect of increased labor can be offset. Moreover, if $L_t$ is increased sufficiently, the productivity effect may become negative; and, in addition, strongly adverse environmental and political effects will be generated. Ultimately, the latter must dominate because, beyond some point, enlargement of the labor input will cause the

[9] See the author and Svetozar Pejovich (1970, pp. 434–37). Change in the variable $C_t$ implies a change in the income or consumption achieved by the original membership of the collective ($L_1$). See equation (10) below.

original majority to lose assured political control of the firm and, hence, to lose control over all of the variables deciding the utility level.

Of course, the obvious strategy to guard against such an outcome is found in restriction of employment. The original majority can be conceived as establishing an *upper limit* on the total size of the working collective:

$$(5) \qquad L_t \leq \hat{L} \quad t=2, 3, \ldots, T$$

As the inequality suggests, the firm's labor input is open to variation from period to period, but in no instance will the number of workers exceed the limit indicated by $\hat{L}$. The parameter $\hat{L}$ reflects the group's judgment on how large a work force can be tolerated if effective political control is to be maintained. The relations (5) and (4) together define the attitudes or preferences of the original majority relative to policy choices for the firm.

## II. The Income Choices Facing the Collective

The given technical, economic, and institutional parameters of the firm impose limits on the rewards the organization can provide for its members. In effect, there exists a set of alternative time streams of real income that is attainable, and the decision makers must choose from among the options in this set. It is important to recognize, however, that the income streams have both pecuniary and nonpecuniary components; thus, something beyond simple monetary accounting is needed. The nature of the choice set can be clarified by considering some specific examples of the rewards that are available. In the simplest case, the decision makers may follow the elementary strategy of doing nothing and merely maintain the firm's initial factor endowment ($L_1$, $K_1$). Then *ceteris paribus*, the workers' pecuniary income will be the same period after period to the planning horizon. This approach avoids any environmental or political changes (nonpecuniary effects) by freezing the voting population at $L_1$, but it also rules out the possibility of the firm achieving a superior income stream. And the latter result is certainly conceivable. By varying $L_t$ or $K_t$, or both, it may be possible for the decision makers to generate pecuniary and nonpecuniary returns that represent a higher level of total utility than that offered by the status quo strategy ($L_1^0$, $K_1^0$).

If the firm's capital stock is to be increased beyond the initial level, a definite plan to finance the expansion has to be established. Logically, three alternatives are open; capital accumulation can be financed entirely by externally supplied credit, entirely by the firm's retained earnings, or by a combination of credit and self-finance. In the interest of simplicity, however, the present model will assume that the self-finance option represents the *only* means by which the firm can provide for capital expansion. In other words, no long-term investment funds are available from the banking system, the central government, or any other external agency. For the capital stock to be enlarged, the workers must elect to reduce the wage fund in some period or periods and "invest" in the firm.

Each investment program considered is necessarily associated with a complementary employment program and, thus, with certain *nonpecuniary* effects (via the $X_{kt}$ variables) that influence total satisfaction. Nevertheless, if the preferences of the original majority are well defined in (4) and (5), an assessment of the net benefits of any investment-employment strategy can always be made. By manipulating $K_t$ and $L_t$, the decision makers are able to establish an ideal balance between pecuniary and nonpecuniary rewards. The result represents the best stream of real income that can be obtained when reliance

is placed *solely on the resources of the firm.*

Unfortunately, however, the choice problem facing the collective is more complex than that of selecting the most desirable real income stream from the set of alternatives made possible by the firm's operations alone. Another basic investment option exists for the workers but this option has its origin *outside the firm.* Even under labor-management, an individual is permitted to save and acquire assets in the form of deposits at a state bank. Since savings deposits yield interest, individual investment in these fully owned funds constitutes a genuine alternative to collective investment in the firm. What emerges, then, is a system where the choice set includes two distinct types of income generating assets. That is, significantly different property rights attach to (owned) savings deposits and (nonowned) capital goods, but the workers have to judge the relative attractions of these two forms of wealth and allocate their savings between the two in optimal fashion.

The latter task is not without its difficulties. This is so because in general it is not possible to make *direct* comparisons between the rate of return on owned assets ($s$) and the rate of return on investment in the firm ($r'$). Differences in property rights account for the problem here. In the case of self-financed investment in capital goods, the principal sum (as such) is never returned to the workers. By contrast, savings banks return both interest and principal. Thus, to assess the true pattern of yields in existence, certain adjustments have to be made. In effect, the return on capital invested in the firm ($r'$) has to be reduced via an appropriate conversion formula to a level ($r$) that workers can compare meaningfully with the return on savings deposits ($s$).

The details of the conversion procedure will be explained below. At this point, it is sufficient to say that correction can be made for the different conditions of ownership and that, depending on the parameters characterizing the labor-managed system, the investment solution reached by the firm's workers will assume one of three possible forms. These are: case (i) Voluntary self-financed investment in the firm is zero and all of the savings recorded by the collective are channeled into deposits at the state banks; case (ii) No funds are allocated to savings deposits but, instead, any income not consumed is devoted to investment in capital goods for the firm; and case (iii) The collective's savings are allocated both to bank deposits and to investment in nonowned capital goods. The precise formal conditions that must be met for each of the respective cases to appear are developed from the general optimization model of the firm and will be presented in Section IV.

With this brief outline of the collective's choice problem in mind, the next step is to examine more closely the different possibilities workers have for securing income over time. The basic stream of revenues comes from the firm itself; and what is needed for the model is an analytical formulation defining the various operating positions permitted by the technical, economic, and institutional characteristics of the production unit. To this end, a so-called "consumption opportunity" locus can be established as follows.

By assumption, the firm's output is produced subject to the stock-flow production function (1). Then, if the firm is competitive and can sell any output at a fixed price $P_t$ set by the government, the profit function emerges as:

$$\Pi_t = P_t \cdot f(L_t, K_t) - Z(K_t) \tag{6}$$

The firm's "profit" in any period ($\Pi_t$) is the difference between total sales revenue and the corresponding "production expenses" indicated by the function $Z(K_t)$.

Production expenses include such outlays as depreciation charges, expenditures on raw materials, interest payments on government-owned capital, etc.; and for convenience the total cost here is taken to be a single valued function of the capital stock.

Under the labor-managed system, the firm's workers are empowered to decide on wage policy. Effectively, this means that the original majority allocates enterprise profit ($\Pi_t$) each period between the wage fund ($S_t$) and the investment fund ($I_t$).[10]

$$(7) \qquad \Pi_t = S_t + I_t, \qquad \Pi_t \geq I_t \geq 0$$

This accounting presupposes that workers are willing to consider the possibility of investing in the firm, and the idea is not unreasonable. While the worker-investors have no transferable rights in any capital accumulated through such *self-finance* activity, they can look forward to higher wages in future periods if the capital increments increase the average productivity of the labor force. In effect, the workers are able to make perpetual investments in the firm and secure "dividends" in the form of enhanced wages. As in the conventional case, investment in what can be termed "nonowned" assets implies an immediate sacrifice because current wages become progressively lower as the allocation $I_t$ increases and the wage fund $S_t$ diminishes.[11]

When the initial employment level is $L_1$ and the initial capital stock is $K_1$, it follows from (6) that the amount of profit is determinate—say, $\Pi_1$. Then, assuming that all of the wage payments go into consumption each period ($S_1 = C_1$), the level of self-financed investment $I_1$ must vary with the desired magnitude of $C_1$:

$$(8) \quad I_1 = P_1 \cdot f(L_1, K_1) - Z(K_1) - C_1$$

[10] Other internal funds are conceivable but are not considered in the model. See the author and Pejovich (1972, pp. 284–87).

[11] See the author (1971, p. 186), Atkinson, pp. 390–91, Vanek (1970, pp. 307–08).

Positive investment has consequences for the firm's position in the next period. If $p$ represents the fixed price per unit of capital equipment, an outlay of $I_1$ dinars translates into a certain number of physical capital units ($I_1/p$) that the firm can add to its existing capital stock ($K_1$) and utilize in the second period. Of course, under long-run conditions, the labor input ($L_t$) may also change along with the capital stock so that the size and composition of the collective changes.[12] The investment situation in period two is analogous to that of the first period. Both the output and total profit of the firm are larger, but self-financed investment depends on the collective's consumption choice ($C_2'$).

$$(9) \qquad I_2 = P_2 \cdot f(L_2, K_1 + I_1/p) - Z(K_1 + I_1/p) - C_2'$$

Here the variable $C_2'$ represents the total consumption of the *new, larger collective*. Its magnitude, however, emerges from the decision of the *original majority* on the ideal size of the wage fund ($S_2 = C_2'$). Of course, following similar reasoning, it is possible to establish the firm's investment equation for each successive period to the planning horizon T. At every stage, the labor force may be varied and a new political constituency may come into existence; the extent of the hiring each period will be determined as part of the solution to the majority's optimization problem. As long as the original majority retains control, both the firm's investment and employment policies will be decided by this dominant group to serve its own interests.

The immediate objective is to derive the

[12] It is assumed here that a competitive firm is able to recruit additional workers with little difficulty. Note, however, that the worker who contemplates joining a new firm must consider not only the current wage but the prospective time stream of pecuniary and nonpecuniary income he is likely to secure by joining the new organization.

consumption opportunity locus *that is perceived by the initial majority of the collective.* This particular construct is necessary because in formulating the optimization model, the variables in the constraint and the variables in the group preference function (4) must be consistent. As a practical matter, the consumption variables $C_2'$, $C_3'$, . . . , $C_T'$ have to be translated into new terms that reflect the options available to the membership of the original collective. First, a purely mechanical adjustment can be made. Assuming that the wage fund is always divided equally among all of the employees of the firm, we have:

$$(10) \qquad C_t = \frac{L_1}{L_t} \cdot C_t', \qquad t = 2, 3, \ldots, T$$

Each period, a proportional share of the firm's wage fund goes to the original members ($L_1$) and is spent by them on consumption as $C_2, C_3, \ldots, C_T$.

With the information supplied by (10), it is possible to rewrite the firm's investment relations. Equation (8) remains the same but all the others change so that the new system becomes the equations shown as (11). *Ceteris paribus*, the volume of self-financed investment planned in any period depends on the contemporaneous volume of consumption by the *original membership* of the collective ($C$); and this volume of consumption is in turn decided by the initial majority. Moreover, assuming that the latter group conforms to rule (5) and restricts employment so as to maintain political control, it follows that $I_T$ must be equal to zero. There is no financial incentive for workers who are about to retire to invest in the firm during period T.[13]

System (11) is so constituted that an investment expression for any period t can be substituted into the succeeding equation relating to period t+1. For example, the expression for $I_1$, involving $L_1$, $K_1$, and $C_1$, can be substituted into the second equation; and, then, the expression for $I_2$ can be introduced into the next equation, etc. As a result of successive operations of this type, the desired opportunity locus is obtainable.

$$(12) \qquad \Psi(C_1, C_2, \ldots, C_T, L_1, L_2, \ldots, L_T) = 0$$

According to our assumptions, $L_1$, $K_1$, and certain market and technical parameters are given; the remaining elements $C_t$ and $L_t$ are variables of direct interest to the decision makers. Operating positions desired by these workers will always lie on the opportunity locus rather than below it, and thus (12) is written as an equality.

## III. The Optimization Problem Restated

For the purposes of exposition, it will be convenient to develop the formal theory of the labor-managed firm in two stages. First, the long-run adjustment process of the firm will be explained on the assumption that decision makers have *only one in-*

[13] Workers might desire to invest in the firm, not for personal gain but for altruistic or patriotic reasons. On balance, however, it does not seem likely that investment motivated in this way will be quantitatively significant.

$$(11) \qquad \begin{aligned}
I_1 &= P_1 \cdot f(L_1, K_1) - Z(K_1) - C_1 \\
I_2 &= P_2 \cdot f(L_2, K_1 + I_1/p) - Z(K_1 + I_1/p) - \frac{L_2}{L_1} C_2 \\
&\vdots \\
I_T = 0 &= P_T \cdot f\left(L_T, K_1 + \sum_{t=1}^{T-1} I_t/p\right) - Z\left(K_1 + \sum_{t=1}^{T-1} I_t/p\right) - \frac{L_T}{L_1} C_T
\end{aligned}$$

*vestment option*. Specifically, the initial optimization model will presuppose that the controlling group chooses, each period, between self-financed investment in the firm and consumption. Then, after the characteristics of this basic construct have been established, the model will be extended to consider the behavior of the firm when savings banks exist and workers have choice among consumption, savings deposits, and self-financed capital investment.

Turning to the first task, the approach is as follows. The key building blocks (4), (5), and (12) suggest that the problem is one of constrained maximization. From the standpoint of the original majority of the collective, the objective is to maximize the multiperiod utility index (4) subject to a series of constraints.[14] Thus:

$$\text{(13)} \qquad \text{Max } U = \phi[C_t, \theta_{kt}(L_t)],$$

$$t = 1, 2, \ldots, T \quad k = 1, 2, \ldots, s$$

subject to $\Psi(C_t, L_t) = 0$,

$$C_1 \leq P_1 \cdot f(L_1, K_1) - Z(K_1) = G_1(L_1)$$

$$\frac{L_2}{L_1} C_2 \leq G_2(C_1, L_1, L_2)$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$\frac{L_{T-1}}{L_1} C_{T-1} \leq G_{T-1}(C_1, C_2, \ldots, C_{T-2}, L_1, L_2, \ldots, L_{T-1})$$

$$L_t - L_1 \geq 0, \qquad t = 2, 3, \ldots, T$$

$$\hat{L} - L_t \geq 0, \qquad t = 1, 2, \ldots, T$$

$$C_t \geq 0, \qquad t = 1, 2, \ldots, T$$

Apart from the fundamental limitation on consumption possibilities imposed by locus (12), certain other restrictions hold because the collective is not permitted to borrow against its expected future income, or to liquidate existing capital stock, in order to increase the wage fund and consumption in any current period.[15] Thus, $C_1$ can be no larger than the firm's total profit in period one, and an analogous limit applies to consumption in each subsequent period. These constraints are obtainable from (11); *ceteris paribus*, the maximum consumption in any period t appears when $I_t = 0$.

Since there are political and other reasons why the labor-managed firm may find it difficult to discharge employees,[16] the model confines attention to the expansion of the firm and assumes that total employment will never fall below the given initial level. The condition $L_t \geq L_1$ holds; but taking account of the upper limit on employment (5), it is also true that $L_t \leq \hat{L}$. Finally, in view of the definitional scheme used in the model, rational interpretation of (13) requires nonnegativity conditions to hold for the variables $C_t$.

Consideration of the mathematical properties of system (13) suggests that the maximization problem can be solved by using the quasi-concave programming methods developed by Kenneth Arrow and Alain Enthoven. Specifically, if the objective and constraint functions in (13) are differentiable and quasi concave, and if the system meets certain other mild requirements, the standard Kuhn-Tucker conditions are both necessary and sufficient for

[14] The model assumes that the majority group possesses certain information about the system in period one, and that on the basis of this information, the group can formulate plans for the optimal adjustment of productive inputs for all remaining periods to the planning horizon. It does not necessarily follow, however, that the plans made in period one will all be carried out. As new information and experience come to the firm over time, the optimization procedure can be repeated. In principle, successive output-investment-employment plans covering periods 2, 3, . . . , T; 3, 4, . . . , T; etc., could be established period after period as the original horizon is approached more closely and more information is available—particularly on political events within the firm.

[15] The restrictions conform to the laws of Yugoslavia.

[16] See the author and Pejovich (1973, pp. 283–86).

a constrained maximum of $\phi$. These technical conditions can be met. The group utility function (2) is ordinal and can be taken as globally strictly quasi concave because it is plausible to assume that the marginal rate of substitution between any pair of variables in the function $\phi(C_t, X_{kt})$ increases monotonically as more of one item is substituted for the other. Ultimately, of course, function (4) appears in the optimization model, but the transformation of (2) into (4) does not change anything essential. Equation (4) is also strictly quasi concave and no difficulty need arise here.[17] The shape of the opportunity locus $\Psi$ as well as the characteristics of the consumption restrictions are determined by the nature of the firm's production function. And when the latter is homogeneous of degree one, all is well. For example, $\Psi$ is clearly concave (and hence quasi concave) if either the labor input or the capital input is held constant while the other is varied. Similarly, if the firm varies both inputs in the same proportion, the opportunity locus will be linear (and quasi concave). The mathematical properties of the constraints become troublesome only when the production function shows increasing returns to scale. But even then reasonable solutions are always possible because of the existence of the upper limit on employment ($\hat{L}$). There are no circumstances in which the firm can pursue sustained expansion subject to increasing returns.

System (13) is formulated so as to emphasize the economic logic of the model and because of this, certain constraints are not set up in standard form for mathematical programming. In what follows, however, the problem will be transformed appropriately—while preserving the quasi concavity of the relations in (13). Then, familiar procedures can be used to obtain the ideal long-run plan for the firm. The first step is to set up the Lagrangian function:

$$
\begin{aligned}
(14)\quad & F(C_t, L_t, \lambda_r) \\
&= \phi[C_t, \theta_{kt}(L_t)] + \lambda_1\Psi(C_t, L_t) \\
&\quad + \lambda_2[-\Psi(C_t, L_t)] \\
&\quad + \lambda_{31}[G_1(L_1) - C_1] + \dots \\
&\quad + \lambda_{3T-1}\left[G_{T-1}(C_1, \dots, C_{T-2}, \right. \\
&\quad \left. L_1, \dots, L_{T-1}) - \frac{L_{T-1}}{L_1} C_{T-1}\right] \\
&\quad + \sum_{t=2}^{T} \lambda_{4t}(L_t - L_1) \\
&\quad + \sum_{t=1}^{T} \lambda_{5t}(\hat{L} - L_t)
\end{aligned}
$$

The relevant Kuhn-Tucker conditions are:

$$
\begin{aligned}
(15)\quad & \frac{\partial F}{\partial C_t} \leq 0, \quad C_t \geq 0, \quad C_t \frac{\partial F}{\partial C_t} = 0 \\
& t = 1, 2, \dots, T \\
& \frac{\partial F}{\partial L_t} \leq 0, \quad L_t \geq 0, \quad L_t \frac{\partial F}{\partial L_t} = 0 \\
& t = 1, 2, \dots, T \\
& \frac{\partial F}{\partial \lambda_r} \geq 0, \quad \lambda_r \geq 0, \quad \lambda_r \frac{\partial F}{\partial \lambda_r} = 0 \\
& r = 1, 2, \dots, 4t, \dots, 5T
\end{aligned}
$$

With the mathematical properties assumed for the model, conditions (15) are both necessary and sufficient for a constrained maximum of $\phi$.[18] A solution is ob-

[17] The indifference contours of function (4) involving $C_t$ and $L_t$ will show positive slopes, but each such contour represents the lower boundary of a convex set and, thus, quasi concavity is present. See Arrow and Enthoven, p. 780.

[18] The constraint set is convex and thus the Kuhn-Tucker constraint qualification holds. Further, the criterion function (4) is defined so that satiation is ruled out. See Arrow and Enthoven, pp. 790–91, and Blaine Roberts and David Schulze, pp. 377–89.

tainable therefore and what it represents is a set of ideal, or utility maximizing, values for the consumption and labor variables—viz., $C_1^*, C_2^*, \ldots, C_T^*, L_1^0, L_2^*, \ldots, L_T^*$. Of course, knowing these, it becomes possible to go back through the system of equations and determine for each period of time to the horizon the optimal levels for: (i) the environmental indexes ($X_{k_t}^*$); (ii) self-financed investment ($I_t^*$); and (iii) output ($q_t^*$).

It should be reemphasized that the respective "optimal" values are optimal from the standpoint of the *original majority* whose preferences are reflected in (4) and (5). Such preferences can take on various configurations and thus the firm's behavior is subject to wide variation. The decision makers may approve self-financed investment and values of $I_t^*$ greater than zero but, on the other hand, a type of *corner solution* can occur if the original majority has great impatience for present consumption and establishes time-indifference manifolds that preclude any sacrifice of current consumption for investment purposes. Similarly, the placement of $\hat{L}$ can allow wide latitude for change in the employment level or constrain the firm to remain at $L_1$. Naturally, there are other parameters of importance in the system beyond those relating to the utility function and the employment limit. The model summarized in (13) is written in highly general notation and this treatment, while saving space, does tend to obscure some significant details. If a more explicit mathematical formulation were given to (13), greater attention could be focused on: (i) the technological coefficients of the production function; (ii) the commodity and factor prices ($P_t, p$) embedded in locus (12) and in the other consumption constraints; (iii) the characteristics of the expense function ($Z$); (iv) the firm's initial factor endowment ($L_1^0, K_1^0$); (v) the internal political conditions of the collective and the size of the original majority group; (vi) the shapes of the conversion relations ($\theta_{kt}$); and (vii) the planning horizon adopted by the decision makers (T).

The particular equilibrium solution reached by the firm depends on the values of all these elements, and any variation in such values must lead to a shift in the firm's position. As a practical matter, the effects of parameter change could be assessed systematically by using comparative statical analysis, but this extension lies beyond the objectives of the present paper. While the significance of changes in the planning horizon for allocative efficiency will be considered in Section IV below, T is the only parameter that is singled out for special study here.

In general, however, the response of the decision makers to policy choices can be analyzed as follows. From (15), it is apparent that at equilibrium the marginal rate of substitution between any two variables in the utility function, as $C_i > 0$, $C_j > 0$, is equal to the ratio of the marginal imputed "costs" associated with these variables. Thus:

$$(16) \quad -\frac{\partial C_j}{\partial C_i} = \frac{\left.\dfrac{\partial U}{\partial C_i}\right|_{C_i^*}}{\left.\dfrac{\partial U}{\partial C_j}\right|_{C_j^*}} = \frac{\sum_{h=1}^{m} \lambda_h^* \left.\dfrac{\partial V_h}{\partial C_i}\right|_{C_i^*}}{\sum_{h=1}^{m} \lambda_h^* \left.\dfrac{\partial V_h}{\partial C_j}\right|_{C_j^*}}$$

where $V_h$, $h=1, 2, \ldots, m$ represent the constraint functions of the system that are *binding* on the variables $C_i^*$, $C_j^*$ at equilibrium, and $\lambda_h^*$, $h=1, 2, \ldots, m$ are the corresponding Lagrangian multipliers. The latter, of course, are all greater than zero. Costs arise in the situation considered because greater consumption in any period $t=i$ implies via constraint (12) some reduction in the consumption levels, and possibly in the environmental quality levels ($X_{kt}$), at other periods coming before or after $t=i$. Whether intertemporal ($\partial C_j/\partial C_i$) or "intercommodity" ($\partial C_t/\partial L_t$)

tradeoffs are considered, the original majority must always balance prospective utilities against opportunity costs. In this way, substitutions proceed until the optimal plan has been formulated and no further advantageous reorganizations can be made.

## IV. Savings Deposits and the Behavior of the Firm

The model discussed in Section III applies to a labor-managed firm operating in an environment where self-financed investment in nonowned assets represents the *only alternative* to consumption. But, as noted earlier, decision makers also have the possibility of owning interest bearing savings deposits; and the existence of this option changes the situation considerably. With investment choice enlarged, the behavior of the firm predicted by system (13) does not necessarily come about. That is, the solution values $C_t^*$, $L_t^*$, $I_t^*$, etc. do not hold unless the magnitudes of certain critical parameters fall into a special pattern. To understand the new (extended) choice problem, it is essential to relate the basic optimization model (13) to the income opportunities offered by the savings banks. When both owned and nonowned assets exist, what takes on central importance for the behavior of the firm is the volume of investment planned in *each asset category*. How this key investment decision is made, and how the investment "portfolio" chosen affects the firm's operations, will be explained in detail below.

One of the fundamental building blocks of model (13) is the consumption opportunity locus (12). The latter defines the set of consumption or income alternatives the firm's resources can provide for the workers. In general (12) indicates that many different time streams of consumption are attainable, but the relation also implies that at any given time period t the next period's choices reduce to:

$$(17) \quad C_{t+1} = H(L_{t+1}, C_t, L_t, \ldots, C_1, L_1)$$

Broadly speaking, the consumption level at t+1 depends on the consumption-investment-employment decisions that have been implemented in the past. At the end of period t, the only variables subject to policy choice are $C_t$ and $L_{t+1}$. Consequently, if $L_{t+1}$ is specified, the variable $C_{t+1}$ becomes a function of $C_t$. In other words, the level of consumption (income) in t+1 has a systematic linkage with the savings and investment behavior of the collective in period t.

Given the past actions of the firm, the maximum consumption level attainable in period t is easily established. This value, say $C_t^0$, results if decision makers refuse to undertake any self-financed investment ($I_t$) in period t. But if $I_t$ is zero, the firm's capital stock does not change. Then, given a value for the labor input $L_{t+1}$, the feasible consumption of the collective ($C_{t+1}^0$) in period t+1 becomes determinate via (6). Of course, as $I_t$ is increased from zero, actual consumption ($C_t$) in period t falls but consumption ($C_{t+1}$) in period t+1 increases. The tradeoff here can be explained in familiar terms. Since it is assumed that $L_{t+1}$ is specified (and since, in any case, an upper limit $\hat{L}$ exists on employment), diminishing returns to capital investment must ultimately appear; in short, the magnitude $C_{t+1}$ must increase at a decreasing rate if the sacrifice of $C_t$ is pushed far enough.

The preceding discussion provides the basis for defining the *average rate of return* ($r'$) on capital invested in the firm. Thus for any two periods t and t+1, we have:

$$(18) \quad r' = \frac{C_{t+1} - C_{t+1}^0}{C_t^0 - C_t} = \frac{C_{t+1} - C_{t+1}^0}{I_t}$$

The consumption level $C_{t+1}^0$ can be achieved when $I_t$ is zero, and it is the difference between this value and the higher

consumption level made possible by positive investment that represents the gain from investment observable in period t+1. The gain in real income is, however, not limited to period t+1. Since the larger capital stock will be sustained indefinitely over time, the increment $(C_{t+1}-C^0_{t+1})$ will appear each period in perpetuity. Equation (18) is then a valid expression for determining the average rate of return.

Under labor-management, the value of the firm's capital stock must be preserved in perpetuity; moreover, the value of any additions to the original stock must also be maintained, regardless of how these additions have been financed. This institutional arrangement has the effect of making self-financed capital accumulation much less appealing in practice than the magnitude of the average rate of return $(r')$ might suggest. Worker-investors do not have permanent claim on the capital stock of the firm and cannot sell their income rights to others when they leave the organization. Thus, to secure appropriate reward, workers must get back both principal and interest payments from the larger stream of income or *wage earnings* investment brings about, and they must do this during the period of time they remain with the firm.

The situation just described means that the effective return on nonowned assets is less than $r'$; correction has to be made for the special property rights structure extant. From the standpoint of actual investment behavior, the relevant rate is determined as follows:

$$I_t = \sum_{\tau=t+1}^{T} \frac{(r' \cdot I_t)_\tau}{(1+r)^\tau} \tag{19}$$

Here, $r$ is the *corrected average return;* it is the average rate of return when adjustment has been made for the length of the collective's planning horizon (T) and for the fact that the principal sum invested $(I_t)$ is never returned as such. Obviously, (19) is nothing more than a present value formula. The expression in the numerator represents the gross increase in income $(C_{t+1}-C^0_{t+1})$ that can be secured *each period* from period t+1 to period T as a result of the self-financed investment $I_t$. Since increments to the firm's capital stock are maintained perpetually, any gain in consumption (income) realized through investment can be sustained to the planning horizon without any further sacrifice by the workers in periods after t.[19] The important thing, though, is that with the aid of (19), any uncorrected rate of return $r'$ can be translated into a corresponding $r$ value that is directly comparable with the return on fully owned assets $(s)$.

For convenience, it will be desirable to rewrite (19) in simpler notation:

$$r = r(I_t, L_{t+1}) \tag{20}$$

The equation presupposes an extensive set of background conditions relating to employment and the firm's capital stock but *ceteris paribus* the average return depends on self-financed investment in period t. Similar expressions for the corrected rate of return could be developed for each period to the planning horizon T. By definition, $rI_t$ gives the total increase in income or consumption $(C_{t+1}-C^0_{t+1})$ that the collective can enjoy per period from period t+1 to period T as a result of an investment $I_t$. The product $rI_t$ represents a *net* increment to income in the sense that full adjustment has been made for the property rights structure. Then, differentiating this product partially with respect to $I_t$ establishes the *corrected marginal return* to capital:

$$\frac{\partial[r(I_t, L_{t+1}) \cdot I_t]}{\partial I_t} = r + I_t \frac{\partial r}{\partial I_t} \tag{21}$$

[19] It is assumed that at each period t+1 through T, the entire increase in the firm's profit $(\Delta\Pi_\tau)$ attributable to the added capital goods $(I_t)$ is paid out in the form of higher wages. Other plausible assumptions about the reward stream are conceivable and can be handled in variant formulations of (19).

The condition derived is important because it shows the change in the magnitude of the (adjusted) consumption increment that occurs in response to an additional small outlay on investment in nonowned assets. Since it is known from the structure of the model that both $r'$ and $r$ must ultimately fall as $I_t$ increases, the product $rI_t$ must ultimately *increase at a decreasing rate* and thus the marginal curve (21) has to fall. That is, in interpreting (21) it should be understood that $\partial r/\partial I_t < 0$.

The expression for the corrected marginal return to investment was obtained directly from model (13) and reflects the productive options open to the labor-managed firm. What has to be done next is to link this condition with the reward that can be secured from fully owned savings deposits. In a system where both owned and nonowned assets are available, utility maximization has to proceed subject to the various constraints specified in (13) and in addition must be consistent with the following requirement:

$$(22)\quad r + I_t \frac{\partial r}{\partial I_t} \geq s, \quad \mathrm{t} = 1, 2, \ldots, \mathrm{T} - 1$$

At any period t, a greater or lesser volume of self-financed investment may be undertaken but, at the margin, the effective productivity of such investment must be greater than or equal to the interest yield on savings deposits ($s$). The rationale for (22) rests on considerations of efficiency; *ceteris paribus* each dollar of savings generated by the collective should be allocated to the most remunerative alternative open. From the standpoint of optimization, additional investment in the firm is justified as long as more capital adds to the total satisfaction (utility) of the decision makers and brings in a return at least as great as the return that would result if the same sum were applied to savings deposits. Of course, once the marginal return to investment sinks below $s$, the collective's funds can be put to better use in banks and investment in nonowned assets should be ended.

When the basic model of the labor-managed firm is modified by the introduction of (22) as an additional constraint, the behavior of the firm is seen in new perspective. Together, systems (13) and (22) can yield three distinct types of solutions:

*Case (i)*: There will be no incentive for workers to invest in the firm if productivity conditions are such that the corrected marginal return (say $m$) is consistently low relative to the interest rate ($s$). That is, self-finance is completely ruled out if $m$ is *less than* $s$ for all positive values of $I_t$ in period one, and in each subsequent period. When $m$ is always less than $s$, workers seeking to maximize utility must: (a) make the size of the wage fund as large as possible consistent with environmental and political concerns, and (b) allocate whatever savings are desired to savings accounts at the state banks. The implications of this type of program for the firm's actions can be summarized as follows:

$$(23)\quad \begin{aligned} I_t &= 0 & \mathrm{t} &= 1, 2, \ldots, \mathrm{T} \\ \overset{0}{L_1}, L_t &= L' & \mathrm{t} &= 2, 3, \ldots, \mathrm{T} \end{aligned}$$

In other words, investment in the firm is nil, but depending on the initial capital-labor ratio and majority preferences (4), some adjustment may be made in the labor input. For example, the labor force may be enlarged in period two (to $L'$) in order to increase the firm's profits and the wage fund. Once the equilibrium adjustment has been achieved, though, no further variation in labor is justified because no change occurs in the capital stock over time. Of course, if the initial conditions do not warrant any departure from the original employment level, $L_1^0$ will be the labor input for all periods to T, i.e., $L' = L_1^0$.

*Case (ii)*: If workers have relatively

limited desire to forego current consumption in order to invest and if the productivity of capital in the firm is high, investment may be confined exclusively to nonowned assets. In such case, no funds will be allocated to savings deposits (owned assets) because in any period when positive investment is undertaken, the corrected marginal return ($m$) remains greater than the rate of interest ($s$). Given the parameters, the constraint imposed by voluntary saving always becomes binding before the return $m$ falls to the level of the interest rate $s$. As far as the behavior of the firm is concerned, the outcome here is essentially like that described by the basic model of Section III, i.e., system (13) without the addition of (22). Practically, the existence of the savings bank option does not influence the solution; the solution values reached ($C_t^*$, $L_t^*$, $I_t^*$, $t=1, 2, \ldots, T$) are independent of the rate of interest $s$ as long as the latter does not rise beyond some critical limit.

Since long-run conditions are assumed, the firm is able to vary both the labor and capital inputs. It follows that the capital productivity schedule may be shifted to the right significantly; thus the possibility is increased for an equilibrium of the type where $m>s$. Specifically, if increasing returns to scale hold, the marginal return to investment in the firm will rise over at least a limited range. Then, unless $s$ is greater than $m$ at the outset, no incentive will exist to place funds in savings deposits within that range. The upward course of $m$ will not continue indefinitely, however, because of the reluctance on the part of the initial majority to increase the labor force. Even under the most favorable circumstances, capital accumulation by the firm has definite limits.

*Case (iii)*: When the model is characterized by a still different set of parameters, workers can find it optimal to invest in both owned and nonowned assets simultaneously. This result will tend to occur in situations where the decision makers have substantial willingness to save and/or where the productivity of capital in the firm is low and falls rapidly. In other words, as $I_t$ is increased in any period, the marginal return $m$ can fall to the level of the interest rate $s$ before the savings constraint becomes binding. Thus, (22) holds in the form of an equality ($m=s$) in some or all periods. Beyond the point at which $m=s$, workers still wishing to save can divert funds to savings deposits and earn interest at rate $s$. Both types of assets, then, are taken simultaneously. It is possible for the firm to increase its capital stock but under the conditions of Case (iii), the decision makers will restrict the investment in nonowned assets relative to what would occur in a system where no savings banks existed.

In each case considered above, the fundamental problem for the collective is to maximize utility function (4) subject to certain constraints, including condition (22). As has been shown, various types of solutions emerge depending in an immediate sense on whether: (i) $s>m$; (ii) $m>s$; or (iii) $m=s$. Regardless of what form the solution takes, however, a definite policy line is defined for the firm, and the nature of the firm's long-run behavior can be predicted.

## V. Allocative Efficiency under Labor Management

The question of what constitutes an efficient pattern of resource allocation within a labor-managed economy is complex and can be argued at different levels.[20] For purposes of the present discussion, however, it will be sufficient to accept the view that allocative efficiency holds when certain familiar marginal rules are met.

[20] See Maurice and Ferguson, pp. 23–27, and Vanek (1969, pp. 1010–11).

In other words, an equilibrium reflecting Pareto optimality is taken as the desideratum, and what has to be decided here is whether the behavior of the labor-managed firm is likely to be consistent with this idealized state. Vanek (1969, p. 1007), for one, argues that Pareto efficiency will prevail provided perfect competition rules and, in particular, that there is free entry for firms employing identical technologies. The conclusions of the model developed above do not support such a presumption, however, and the reasons why allocative difficulties can be expected are easily demonstrated.

The literature has shown that in the case of a firm operating over time subject to a stock-flow production function, intertemporal efficiency requires only myopic vision on the part of decision makers. Specifically, an optimal allocation program for the competitive firm will result if, at each cross-section of time, resources are employed in such way as to equate the marginal rate of substitution between any pair of inputs ($\partial K_t/\partial L_t$) with the ratio of their effective prices (rental rates) in that period.[21] Then, considering the simplest possible situation where the initially given product and factor prices are expected to hold unchanged to the planning horizon, it is clear that the firm's allocation pattern reduces to one that is frozen over time; each period, the unchanged price of a factor equals the value of its marginal product, and input flows must be the same in all periods through T. Under the latter conditions, it might be thought that the labor-managed firm could reach the ideal, or Pareto optimal, level of *capital stock* without difficulty if only the equilibrium rate of interest in the economy were known. That is, knowing the interest rate, the firm could in theory follow a standard decision rule and continue to add to its initial capital stock up to the point where the marginal return to capital invested in the firm was equal to the equilibrium interest charge. But, actually, the labor-managed firm will not act so as to attain this ideal position. In general, the institutional structure militates against the possibility of the firm achieving an optimal capital stock.

[21] See Agnar Sandmo, pp. 1336–42.

One obvious trouble arises from the fact that the equilibrium rate of interest is not known. In a labor-managed socialist system, there is no free capital market and hence no practical way of establishing an unambiguous measure of the general opportunity cost of capital. This situation poses a serious problem for efficient allocation. Nevertheless, in order to place the labor-managed system in the best possible light, let us assume away the obstacle and accept the fiction that the equilibrium rate of interest ($s$) is known and *is equal to the rate paid on savings at the state banks.* As explained earlier, the firm's decision makers have a choice between investment in savings deposits and investment in the firm, i.e., between owned and nonowned assets. What has to be emphasized, however, is that the property rights arrangements in existence strongly affect the *relative attractiveness* of these two types of assets. And relative attractiveness is important, of course, because the decision taken on the type and amount of assets to be acquired exerts a direct effect on the input-output behavior of the firm over time.

Nonowned assets have lesser appeal than fully owned savings deposits for the simple reason that the property rights workers possess in nonowned physical capital are limited to *usus fructus.* In practical terms, this condition means that the firm has lesser incentive for capital accumulation than is desirable from a social point of view. From society's standpoint, the marginal return on an increment of

capital added to the firm is given by the relation:

$$\frac{\partial(r'I_t)}{\partial I_t} = r' + I_t \frac{\partial r'}{\partial I_t} = m' \tag{24}$$

where $r'$ is the average rate of return on capital (18) *uncorrected for workers' property rights*. Efficient allocation requires that the uncorrected marginal return ($m'$) be equal to the rate of interest ($s$) at equilibrium. But, as has been shown in Section IV, decision makers at the level of the firm are motivated to invest in nonowned assets only when the *corrected* marginal return ($m$) is greater than or equal to the interest rate ($s$). Obviously, the divergence between the needs of society and the needs of the firm must lead to difficulties.

If Case (i) holds and the corrected marginal return $m$ is always less than $s$, the firm's workers will commit all of their voluntary savings to owned assets. Self-financed investment will be zero and the firm will operate in accordance with the input plan described in (23). This choice, however, is consistent with a misallocation of resources because $m'$ can be greater than $s$ even though $m$ is less than $s$. In other words, it is possible for socially desirable investment in the firm to be neglected despite the fact that the true return on capital is greater than its opportunity cost ($m'>s$). The fundamental reason for the problem is revealed by equation (19). The average rate of return $r'$ must be greater than the corrected rate $r$ for every level of $I_t$ as long as the planning horizon T is finite. What represents a *perpetual* income stream from the standpoint of society ($r'I_t$ per period) is seen as a *finite* stream (to T) by the collective and thus $r'>r$. Of course, the distinction between the two marginal returns follows *mutatis mutandis*, and $m'$ is always greater than $m$. It is also clear that the extent of the difference between $m'$ and $m$ (for any $I_t$) becomes larger as the collective's planning horizon T is shortened.

Positive self-financed investment to increase the firm's capital stock is feasible provided $m \geq s$. The precise outcome here will be given by the solution values of system (13) and (22). While such investment in the firm may be desirable, it remains true that the condition $m \geq s$ *precludes allocative efficiency*. Pareto efficiency requires that investment be carried to the point where $m'=s$; but there is no chance for such a result to occur through worker investment in the firm. At best, Case (iii) holds and the firm's equilibrium will be reached where $m=s$. However, since $m'>m$, it follows that $m'>s$ and the process of self-financed capital accumulation must always be *relatively restricted*. If the firm does not commence its operations with the optimal capital stock, the optimal stock will not be reached.

Is there any escape from the difficulty just noted? It can be argued, of course, that the rate of interest offered on savings deposits does not have to be set at $s$. Individual savers could be confronted with a much lower rate (say, $s_0$) than the equilibrium rate of interest—even though this policy ($s_0<s$) might tend to reduce the volume of savings deposits. In any event, when a low value is established for $s_0$, and when the collective has a sufficiently long planning horizon so that the difference between $m'$ and $m$ is relatively small at each level of $I_t$, the possibility exists for resolving the conflict between the objectives of society and the firm. The separate conditions $m=s_0$ and $m'=s$ *may be met simultaneously*. The collective can find it advantageous to invest in nonowned assets up to the point where $m=s_0$; but given the right parameters optimal capital accumulation will also take place because $m'=s$ when $m=s_0$.

In a system where $m'>m$ and $s>s_0$,

manipulation of the $s_0$ value can be useful. There is, however, no assurance that the planning horizon chosen by the firm's decision makers will always be long enough to permit a sufficient reduction of $s_0$ and the possibility of $m'=s$ while $m=s_0$. Moreover, self-financed investment can easily be zero despite a low value for $s_0$ and a favorable value for T because other parameters are involved in the determination of investment policy. Ultimately, the firm's decision makers are concerned with the effect of investment on the utility index (4). This implies that nonpecuniary as well as pecuniary considerations affect the firm's choice of inputs. Thus, in so far as investment is tied to *expansion of the labor force*, the decision makers may anticipate political and environmental difficulties and vote to reject even those capital projects promising substantial pecuniary returns.

A striking feature of the labor-managed system is the restricted range of investment alternatives it allows the workers. Investment in education and training aside, an individual can choose to accumulate savings deposits, or to invest in his firm, or to do both simultaneously. This arrangement is unsatisfactory, however, because it means that the allocation of capital has to take place subject to inadequate information about other alternatives in the economy. Just as there can be too little investment by the firm, there can be too much if the bank rate $s_0$ is set significantly below the equilibrium rate of interest $s$. That is, at the level of investment $I_t$ where $m=s_0$, $m'$ can be less than $s$; and the result is that the volume of self-financed investment in the firm will then be too large. Inefficient allocation occurs here because the firm's decision makers are not aware of the true value of $s$ and, in any case, they are guided by the relative magnitudes of $s_0$ and $m$.

In general, as long as the corrected marginal return on capital is greater than or equal to the established interest rate ($m \geq s_0$), the investment solution will be made on the basis of the collective's subjective estimates of costs and benefits. The first-order conditions emerging from the solution of model (13) and (22) will decide the outcome independently of circumstances in the rest of the system. For example, assuming the collective's propensity to save is small, Case (ii) is likely to appear and the latter indicates that not even the equilibrium $m=s_0$ will be reached, to say nothing of $m'=s$. As far as the firm is concerned, then, the realization of an optimal capital stock is merely a possibility, a result that can come about only under very special conditions. At any moment of time, the labor-managed firm is controlled by particular individuals having the power to shape the firm's investment, employment, and output policies *to suit their own specialized objectives*. If the decision makers are rational and truly free to seek their own interests, it seems highly improbable that they will behave in the manner suggested by the simple wage maximization hypothesis.[22] In what is likely to be the typical situation, the firm's planning horizon will be relatively short, the wage rate will not be maximized each period, and factor allocation will not conform to the requirements for Pareto optimality.[23]

As explained, the original majority determining the firm's policy will normally find it desirable to limit the input of capital to levels below those required by orthodox

[22] See Ward and Vanek (1970).

[23] Note that the firm's *initial conditions* exert a significant influence on the nature and extent of the productive adjustments the collective will ultimately make. *Ceteris paribus*, the initial capital-labor ratio ($K_1/L_1$) determines the marginal productivity of capital and affects the incentives for subsequent self-financed investment. Similarly, the initially given political structure within the firm influences $\hat{L}$ and decides how far the original majority will be willing to proceed with the hiring of new employees.

marginal productivity considerations. But deliberate restriction of the firm's labor input is also to be expected. Because of the pressure of internal politics and concern for the firm's environment, there will be reluctance on the part of the original majority to increase the labor force. A simple illustration can be used to explain the factors affecting employment policy. Assume that the firm's production function is homogeneous of degree one and that the initial capital-labor ratio $(K_1/L_1)$ is *larger* than the equilibrium ratio required to maximize the wage per worker in the first period. Under these conditions, the employment of additional labor could raise the prevailing wage rate step by step until the maximum wage $(w_1^*)$ was reached. But from the standpoint of the original majority, the desirability of greater employment cannot be judged on the basis of productivity alone. As the first-order conditions (16) indicate, increased labor reduces utility via the effect on the $X_{st}$ variables, and influences subjective costs in so far as the employment constraint $\hat{L}$ becomes binding. Thus, employment may not be carried as far as the usual marginal logic $(P_1\partial q_1/\partial L_1 = w_1^*)$ would demand. Such restrictive behavior is quite rational from the standpoint of the decision makers, but its consequences for efficient resource allocation are predictably bad.[24]

It may seem anomalous that rational individuals would deliberately choose to operate with factor combinations that promise lesser output per worker rather than with alternatives having greater productivity. But as emphasized throughout the analysis, the choices made by the initial majority of the collective are perfectly understandable when the role of *appropriability* is recognized. The rewards that are meaningful to these workers are the pecuniary and nonpecuniary returns that they can actually capture during their period of tenure with the firm. Having neither permanent nor transferable claims on the assets of the labor-managed firm, decision makers operate with relatively short planning horizons and view economic choices in a somewhat distorted perspective. Policies that increase the wealth of the firm over some stretch of time, but do not improve the appropriable real incomes of the decision makers during this same period, are not likely to be considered desirable, and will tend to be rejected. Given these fundamental characteristics, the ultimate conclusion on the labor-managed firm is clear. Whatever its contribution to industrial democracy, it is not an inherently efficient economic organization.

[24] It can be noted here that the problems of allocative inefficiency remain to plague the firm whether or not the firm is able to draw upon bank credit or other external capital funds. See the author and Pejovich (1973, pp. 280–86) and the author (1974).

## REFERENCES


K. J. **Arrow** and A. C. **Enthoven**, "Quasi-Concave Programming," *Econometrica*, Oct. 1961, *29*, 779–800.

A. B. **Atkinson**, "Worker Management and the Modern Industrial Enterprise," *Quart. J. Econ.*, Aug. 1973, *87*, 375–92.

A. **Bergson**, "Market Socialism Revisited," *J. Polit. Econ.*, Oct. 1967, *75*, 432–42.

E. **Domar**, "The Soviet Collective Farm as a Producer Cooperative," *Amer. Econ. Rev.*, Sept. 1966, *56*, 734–57.

E. G. **Furubotn**, "Toward A Dynamic Model of the Yugoslav Firm," *Can. J. Econ.*, May 1971, *4*, 182–97.

———, "Bank Credit and the Labor-Managed Firm: The Yugoslav Case," in his and S. Pejovich, eds., *The Economics of Property Rights*, Cambridge, Mass. 1974.

——— and S. **Pejovich**, "Property Rights and the Behavior of the Firm in a Socialist State: The Example of Yugoslavia," *Zeit. für Nationalökon.*, Band *30*, Heft 3–4, 1970, 431–54.

——— and ———, "The Formation and Distribution of Net Product and the Behavior of the Yugoslav Firm," *Jahrbuch der Wirtschaft Osteuropas*, Band *3*, 1972, 265–87.

——— and ———, "Property Rights, Economic Decentralization, and the Evolution of the Yugoslav Firm, 1965–1972," *J. Law Econ.*, Oct. 1973, *16*, 275–302.

B. Horvat, "Yugoslav Economic Policy in the Post-War Period: Problems, Ideas, Institutional Developments," *Amer. Econ. Rev.*, June 1971, Supp., *61*, 69–169.

S. C. Maurice and C. E. Ferguson, "Factor Usage by a Labor-Managed Firm in a Socialist Economy," *Economica*, Feb. 1972, *39*, 18–31.

R. Radner, *Notes on the Theory of Economic Planning*, Athens 1963.

B. Roberts and D. Schulze, *Modern Mathematics and Economic Analysis*, New York 1973.

A. Sandmo, "Investment and the Rate of Interest," *J. Polit. Econ.*, Nov./Dec. 1971, *79*, 1335–45.

G. Tullock, "The General Irrelevance of the General Impossibility Theorem," *Quart. J. Econ.*, May 1967, *81*, 256–70.

J. Vanek, "Decentralization Under Workers' Management: A Theoretical Appraisal," *Amer. Econ. Rev.*, Dec. 1969, *59*, 1006–14.

———, *The General Theory of Labor-Managed Market Economies*, Ithaca 1970.

B. Ward, "The Firm in Illyria: Market Syndicalism," *Amer. Econ. Rev.*, Sept. 1958, *48*, 566–89.

# Inflationary Expectations: Their Formation and Interest Rate Effects

*By* KAJAL LAHIRI*

In estimating the effects of inflationary expectations on nominal interest rates, distributed lags on past prices have generally been used as observable proxies for the supposedly unobservable price change expectations. Thus, Irving Fisher's famous explanation of the Gibson paradox is expressed in terms of two hypothesized relationships:

$$(1) \qquad rn_t = rr_t + \dot{p}_t^e$$

$$(2) \qquad \dot{p}_t^e = \sum_i w_i \dot{p}_{t-i}$$

The first equation states that the nominal interest rate ($rn_t$) for a particular asset with returns fixed in money terms is equal to the real rate of interest ($rr_t$) plus the anticipated rate of inflation ($\dot{p}_t^e$). He further assumed that $rr_t = \rho + u_t$ where $\rho$ is the long-run real equilibrium rate of interest and where $u_t$ is a random error term. Since equation (2) was used as a proxy for $\dot{p}_t^e$ in equation (1) the reduced form becomes

$$(3) \qquad rn_t = \rho + \sum_i w_i \dot{p}_{t-i} + u_t$$

and the whole problem finally boils down to that of finding out a particular lag structure which best explains the interest rate variations. Robert Gordon followed exactly the same approach to develop an independent measure of price expectations which was used for estimating wage-price equations.

It should be pointed out that the distributed lag coefficients of equation (2) can only be recovered from the reduced form estimates if the coefficient of price expectations in the interest rate equation is assumed to be unity; or some other assumption is made on $w_i$. As William Gibson has pointed out, there are important theoretical and empirical studies which suggest that the nominal rate of interest may not fully adjust to changes in price expectations. Thus, equation (1) should be modified as

$$(4) \qquad rn_t = \alpha + \beta \dot{p}_t^e + u_t$$

and ideally $\beta$ should be subject to usual statistical tests. In the context of equation (3) another assumption which identifies the parameter $\beta$ and which seems to be implicit in a number of studies is that $w_i$ sums to one. This restriction means that eventually people will catch on and fully anticipate inflation. It is certainly reasonable to expect that a sustained constant inflation sooner or later would be fully anticipated; but this consideration is of very little use in producing a reasonable restriction to impose on $w_i$ for the purpose of empirical estimation because during the periods used in estimation, inflation was never so sustained (see Thomas Sargent). Another problem associated with the distributed lag proxy is that it is modelled in a deterministic form which may be quite restrictive in diverse periods. More prop-

* Assistant professor of economics, Indian Institute of Management, Calcutta, and State University of New York, Albany. This paper is taken from my doctoral dissertation. I am especially indebted to my adviser, G. S. Maddala, and to K. M. Gaver, D. J. O'Hara, Sherwin Rosen, and an anonymous referee for helpful comments. Responsibility for any errors is solely mine. Financial support from the National Science Foundation under grant GS-33356 to the University of Rochester is gratefully acknowledged. A special note of thanks is due to William Gibson for kindly supplying the data on observed price expectations.

erly, expectations should be modelled with an error:

$$\dot{p}_t^e = \sum_i w_i \dot{p}_{t-1} + \eta_t \tag{5}$$

Unfortunately modelling expectations stochastically results in a reduced form in which the explanatory variables are correlated with the composite error term. Hence ordinary least squares would give inconsistent estimates.

To overcome these difficulties Gibson and D. H. Pyle used the Livingston survey data (from Philadelphia Bulletin; for more information, see Gibson) on price expectations to test the Fisher effect. In some instances they found the value of $\beta$ to be far below unity. Stephen Turnovsky and Michael Wachter had noted in the context of their study of the Phillips curve that the coefficient might have been seriously underestimated due to the presence of some errors in the measurement of the variable. Gibson and Pyle, too, pointed out the same.

Turnovsky used the same data to study the lag structure of past prices that best explains the data on price expectations. The purpose of the present paper is to estimate the following structural model:

$$\dot{p}_t^* = \dot{p}_t^e + u_{1t} \tag{6}$$

$$rn_t = \alpha + \beta \dot{p}_t^e + u_{2t} \tag{7}$$

$$\dot{p}_t^e = \sum_{i=1}^{m} w_i \dot{p}_{t-i} + u_{3t} \tag{8}$$

where $\dot{p}_t^*$ = observed price expectations
$\dot{p}_t^e$ = unobservable price expectation variable
$\dot{p}_t$ = actual rate of change in prices
and $u_1, u_2, u_3$ are random error terms

Apart from testing Fisher's two hypotheses independently and in a unified framework, the model would take into account one of the simplest types of errors that might have blurred the observed price expectation variable. In the process we also get much more precise estimates of the lag structure than those obtained by Turnovsky (1970).

In two recent papers Jacob Frenkel (1975a,b) has presented a theory of the formation of short-term expectations which combines both regressive and adaptive elements. I have tested the hypothesis against the available data and incorporated it in the above structural model.

The plan of this paper is as follows: Section I contains some statistical preliminaries and some notes on the formation of expectations. Section II analyzes the results and Section III discusses a possible break in the structural relations around 1960. Finally, concluding remarks are given in Section IV.

## I. Data and the Formation of Expectations

From Livingston's survey two series of expected prices can be obtained—one from six months ahead (short term) and the other from twelve months ahead (long term) predictions of the price level. Since both Gibson and Pyle found the coefficients of the 6-month price expectations to be particularly low, this study has been restricted to the estimation of Fisher effect with short-term price expectations only. Data on market yields on 6-month bills were not available prior to 1959; hence yields on 3-month bills were used for the period 1952–70. The 6-month yield observations are monthly averages for June and December.

The basic time interval is one quarter. The following notations are used:

$p_6^*(t)$ = expected percentage price change at time t for the next six months
$p_{12}^*(t)$ = expected percentage price change at time t for the next twelve months
$\dot{p}_t'$ = actual percentage price change

for quarter immediately preceding time t

$\dot{p}_t$ = actual percentage price change for the six months immediately preceding time t
$= \frac{1}{2}(\dot{p}'_t + \dot{p}'_{t-1})$

All quantities are expressed as annual rates and past price changes are based on the Consumer Price Index.

Turnovsky tested three expectation hypotheses: unrestricted distributed lag, adaptive, and extrapolative. For the short-term expectations the equations are:

$$(9) \qquad \overset{*}{p}_6(t) = \sum_{i=0}^{\infty} w_i p_{t-i}$$

$$(10) \qquad \overset{*}{p}_6(t) = w_0 + w_1 \overset{*}{p}_6(t-2) + w_2 \dot{p}_t$$

$$(11) \qquad \overset{*}{p}_6(t) = w_0 + w_1 \dot{p}_t + w_2(\dot{p}_t - \dot{p}_{t-2})$$

The coefficient $w_1 + w_2 = 1$ and $w_0 = 0$ in equation (10) would give the original version of adaptive expectations. On the other hand, the original version of extrapolative hypothesis stipulated $w_0 = 0$ and $w_1 = 1$ in equation (11). The coefficient $w_2 < 1$ in (10) implies expectations to be regressive.

Frenkel (1975a,b) has developed a unified theory of the formation of expectations which tries to explain the impact of money supply on nominal interest rates in the short run as well as in the intermediate and the long run. Specifically he hypothesized that short-term expectations are influenced by two factors: (i) a regressive element, which is determined by the discrepancy between the actual rate of inflation $\dot{p}_t$ and an average long-term expected rate; and (ii) the usual adaptive element which is determined by the discrepancy between $\dot{p}_t$ and the expected short-term rate.

Frenkel gives a numerical example to explain the first element: consider a steady state in which all expectations are realized, and the actual and expected rates of inflation are 10 percent per unit of time. Now, consider a sudden jump in the rate of inflation to 15 percent per unit of time. Since individuals believe on the basis of past history that the average normal rate of inflation is 10 percent, they will expect the rate of inflation for the next unit of time to be below 10 percent, say, 5 percent, so as to average out to the normal mean rate of 10 percent and thus to conform with their expectations on future price level. They might, in addition, revise their expected path of prices and thus their average long-term expectation slightly upwards, in which case they will expect the rate of inflation for the immediate next period to be above 5 percent but still below 10 percent. Regarding the second element, it is assumed that whenever the actual rate of inflation exceeds the expected short-term rate, individuals will revise their short-term expectations upwards by some constant multiple of the discrepancy.

To test the above hypothesis we estimated equations of the following form:

$$(12) \qquad \overset{*}{p}_6(t) = w_0 + w_1 \overset{*}{p}_{12}(t) + w_2 \overset{*}{p}_6(t-2) + w_3 \dot{p}_t$$

The coefficient $w_1$ gives the coefficient corresponding to the regressive element and $(1 - w_2)$ would give the coefficient corresponding to the adaptive element. Thus Frenkel's hypothesis can be tested by introducing long-term expectations as an additional variable in the adaptive hypothesis given in equation (10).

## II. The Results

Ordinary least squares regression of $rn_t$ on $p_6^*(t)$ for the period 1952–70 produced the following results:

$$rn_t = \underset{(.182)}{2.347} + \underset{(.063)}{0.586} \overset{*}{p}_6(t)$$

$$D\text{-}W = 1.449 \qquad \bar{R}^2 = 0.702$$

TABLE 1—ORDINARY LEAST SQUARES TO EXPLAIN THE FORMATION OF EXPECTATIONS 1952–70

| Expectations | $\bar{R}^2$ | D-W |
|---|---|---|
| Weighted Expectations: | | |
| $p_6^*(t)=.251\dot{p}_t+.102\dot{p}_{t-1}+.543\dot{p}_{t-2}$ | .587 | 1.018 |
| (.361) (.389) (.195) | | |
| Adaptive Expectations: | | |
| $p_6^*(t)=.426\dot{p}_t+.534p_6^*(t-2)$ | .580 | 1.989 |
| (.145) (.164) | | |
| Extrapolative Expectations: | | |
| $p_6^*(t)=-.494+1.011\dot{p}_t-.568(\dot{p}_t-\dot{p}_{t-2})$ | .555 | 1.354 |
| (419) (.148) (.183) | | |
| Frenkel's Hypothesis: | | |
| $p_6^*(t)=.786p_{12}^*(t)+.046\dot{p}_t+.535p_6^*(t-2)$ | .668 | 1.962 |
| (.251) (.177) (.145) | | |

*Note*: Standard errors are in parentheses; $\bar{R}^2$ is adjusted $R^2$, and *D-W* is the Durbin-Watson statistic.

Regression results of the four hypotheses to explain the formation of expectations are given in Table 1. The results are very similar to those obtained by Turnovsky though in this case the unrestricted distributed lag model is slightly better than the adaptive and the extrapolative models on the $\bar{R}^2$ criterion. In almost all the cases the constants were found to be insignificant and were accordingly dropped. Since the coefficient of $\dot{p}_t$ in the extrapolative case and the sum of the coefficients of $\dot{p}_t$ and $\dot{p}_6(t-2)$ in the adaptive case were very close to unity, the original versions of these models were found to be successful in explaining the data. Apart from that, a negative coefficient value for $(\dot{p}_t-\dot{p}_{t-2})$ suggests strong regressivity. The introduction of $p_{12}^*$ in equation (12) was found to be a very significant additional variable. The $\bar{R}^2$ was significantly higher and the value of *D-W* statistic was very close to 2. If $p_{12}^*(t)$ can roughly be assumed to represent the (long-term) expected average rate of inflation, then the significance of the coefficient for $p_{12}^*(t)$ implies regressivity corresponding to longer-term expectations. The coefficient of the lagged short-term expectation retained the same value as in the pure adaptive model. The low value for the coefficient of $\dot{p}_t$ is only expected since it really measures the difference between the coefficients corresponding to short-term and long-term adjustments. Thus, it seems that Frenkel's (1975a,b) hypothesis best explains the observed data on price expectations.

The reduced form of the structural model (6)–(8) is

$$p_6^*(t) = \sum_{k=1}^{k} w_k\dot{p}_{t-k} + u_{1t} + u_{3t} \tag{13}$$

$$rn_t = \alpha + \beta \sum_{k=1}^{k} w_k\dot{p}_{t-k} + u_{2t} + \beta u_{3t} \tag{14}$$

This can also be written in the following alternative structural form:

$$p_6^*(t) = \sum_{k=1}^{k} w_k\dot{p}_{t-k} + u_{1t} + u_{3t} \tag{15}$$

$$rn_t = \alpha + \beta p_6^*(t) + u_{2t} - \beta u_{1t} \tag{16}$$

Let $u_{1t}+u_3t=v_{1t}$ and $u_{2t}-\beta u_{1t}=v_{2t}$. Gibson and Pyle estimated the equation (16) by ordinary least squares (*OLS*) where $p_6^*(t)$ is really endogenous. Turnovsky estimated equations of the form (15). Though measurement error in the dependent variable does not lead to any bias in *OLS* estimates, the presence of the composite disturbance term $v_{1t}$ may unnecessarily produce large standard errors of estimates. This would

TABLE 2—TWO-STAGE LEAST SQUARES ESTIMATES OF THE INTEREST RATE EQUATION

| Expectations Hypotheses | Estimates | $\bar{R}^2$ |
|---|---|---|
| Weighted: | $rn_t=1.598+.906\hat{p}_6^*(t)$ | .778 |
| | (.217) (.075) | |
| Adaptive: | $rn_t=2.044+.765\hat{p}_6^*(t)$ | .729 |
| | (.205) (.080) | |
| Extrapolative: | $rn_t=1.976^a+.796\hat{p}_6^*(t)$ | .752 |
| | (.238) (.096) | |
| Frenkel's: | $rn_t=2.077+.723\hat{p}_6^*(t)$ | .704 |
| | (.262) (.080) | |

[a] This is an estimate of $\alpha+\beta w_0$.

TABLE 3—THREE-STAGE LEAST SQUARES ESTIMATES OF THE STRUCTURAL PARAMETERS[a]
1952–70

| Expectations Hypotheses | $w_0$[b] | $w_1$ | $w_2$ | $w_3$ | $\alpha$ | $\beta$ |
|---|---|---|---|---|---|---|
| Weighted | — | .482 | −.038 | .442 | 1.638 | .892 |
| | | (.207) | (.221) | (.116) | (.149) | (.056) |
| Adaptive | — | .487 | .469 | | | .771 |
| | | (.112) | (.122) | | | (.058) |
| Extrapolative | −.510 | 1.018 | −.558 | | 1.984[c] | .791 |
| | (.367) | (.119) | (.126) | | (.193) | (.052) |
| Frenkel's | — | .510 | −.558 | .509 | | .723 |
| | | (.207) | (.126) | (.122) | | (.062) |

[a] Standard errors are in parentheses.
[b] The constant term $w_0$ was estimated only in the extrapolative case; in other cases constants were found to be insignificant and were accordingly dropped.
[c] This is an estimate of $\alpha+\beta w_0$.

particularly be the case if $\dot{p}_{t-i}$ are correlated.

To obtain consistent estimates of $\alpha$ and $\beta$, the two-stage least squares (*2SLS*) procedure was used. The results are given in Table 2 where it is found that the estimated value of $\beta$ has increased from .586 to as high as .906 in some cases. As expected, some loss in efficiency is noticeable since the standard errors have increased uniformly. Since a consistent estimate of of the variance-covariance matrix of residuals showed rather high negative value for the covariance term, I computed the "full information" estimates. These are reported in Table 3.

The standard errors for all the coefficients have decreased considerably. The three-stage least squares (*3SLS*) estimates of $\beta$, though on the average little less than the corresponding *2SLS* estimates, were appreciably higher than the *OLS* estimates. By comparing Tables 1 and 3, it can be seen that much more precise estimates of the lag structures in the formation of price expectations have been obtained.[1]

[1] Since with twelve-month ahead predictions the estimated Fisher effects were quite sensible, I did not introduce any possible error in $\dot{p}^*_{12}$. Also, I assumed that the decision makers respond according to $\dot{p}^*_6(t-2)$ directly. In short, I did not introduce possible errors in the variables $\dot{p}^*_6(t-2)$ and $\dot{p}^*_{12}(t)$ in equations (14) and (16), respectively.

It may be noted that Gordon's approach to develop an independent measure of the lag structure was to estimate equation (14) of the reduced form and to pick up the estimated coefficients of the lagged prices. By *OLS* we obtain the following estimate of equation (14).

$$\underset{}{rn_t} = \underset{(.202)}{1.562} + \underset{(.187)}{.467}\dot{p}_t - \underset{(.198)}{.049}\dot{p}_{t-1} + \underset{(.103)}{.406}\dot{p}_{t-2}$$

Since the $var(u_2+bu_3)$ was 0.56, which is far less than $var(u_1+u_3)$, (which is 2.22), and since the true value of $\beta$ is very close to unity, the lag structure developed in this way is much more precise than that obtained from equation (13). These estimates are strikingly similar to those obtained by *3SLS*.

The estimates of error variances of the original structural model (6)–(8) are given in Table 4. Given the estimates of $var(v_1)$,

TABLE 4—ESTIMATES OF THE ERROR VARIANCES OF THE STRUCTURAL MODEL
1952–70[a]

| Expectations | $V(u_1)$ | $V(u_2)$ | $V(u_3)$ |
|---|---|---|---|
| Weighted | 1.561 | .032 | .656 |
| Adaptive | 1.393 | .166 | .871 |
| Extrapolative | 1.435 | .082 | .897 |
| Frenkel's | 1.152 | .304 | .704 |

[a] Under additional assumption that $cov(u_i, u_j)=0$, $i,j=1,2,3; i\neq j$.

$var(v_2)$, and $cov(v_1, v_2)$, one can recover only three second-order moments of the structural disturbances $(u_1, u_2, u_3)$. The three variances viz. $var(u_1)$, $var(u_2)$, and $var(u_3)$ have been estimated under additional assumptions that $cov(u_i, u_j)=0$ for $i \neq j$, $i, j, = 1, 2, 3$. Though there is no way to test these assumptions, there is some indirect evidence which suggests that the assumptions may be roughly true. Note that the assumptions

$$cov\ (u_i, u_j) = 0 \qquad i, j = 1, 2, 3, \quad i \neq j$$

would imply the following three independent restrictions:

$$(17) \qquad var\ (v_1) = E[v_1^2(t)] = \sigma_1^2 + \sigma_3^2$$

$$(18) \qquad var\ (v_2) = E[v_2^2(t)] = \sigma_2^2 + \hat{\beta}^2 \sigma_1^2$$

$$(19) \quad cov\ (v_1, v_2) = E[v_1(t), v_2(t)] = -\hat{\beta}\sigma_1^2$$

Thus, $cov(v_1, v_2)$ must have a sign opposite to that of $\beta$. Since in our case $\beta$ is always positive, equation (19) implies that $cov(v_1, v_2)$ assumes negative values. This, then, would give a positive $\sigma_1^2$. Similarly, given that $\sigma_2^2$ and $\sigma_3^2$ are positive quantities we get two additional restrictions. In all the experiments these restrictions were always satisfied. In addition, the reduced form covariance $\beta_1\ var(u_3)$ was always positive.

To the extent we can rely on these estimates, the calculations suggest that $var(u_1)$ is quite large, followed by the disturbance term in the expectation formation equation. In all the cases the error variance of the interest rate equation was found to be quite low.

### III. Structural Change Around 1960

The structural model (6)–(8) is particularly suitable for an analysis of possible changes in one or more of the structural relations. William Yohe and Denis Karnosky when relating nominal interest rates to distributed lag of past prices found enough evidence to support a break in the sample around 1960. They attributed the break to a shift in the interest rate equation. But, as Gibson has pointed out, there is also the other possibility that the shift had occurred in the formation of price expectations equation. In fact, the latter possibility has been supported by the experiments reported by Turnovsky.

Gibson found a break in the context of the equation (16); Turnovsky found the same in the contxt of equation (15). But it is interesting to note that the observed break may be due to a break in relation (6). Also, if the reduced form relation (14) does not show any break in the sample, then this finding should suggest that the shifts in relation (15) and (16) are due to shift in the relation (6), and not due to shifts in relations (7) and (8). I tested the propositions with the help of all the four expectations-formation hypotheses (9)–(12). When "extrapolative" and "distributed-lag" hypotheses were used, the Chow tests for stability of the coefficients were resoundingly rejected against the alternative that there was a break by the end of 1959. However, with "adaptive expectations," the equality of regressions hypothesis was accepted at both the 1 and 5 percent level of significance, while with Frenkel's hypothesis it was accepted only at 1 percent level. Thus, it seems that the adaptive expectations hypothesis is quite successful in explaining the reduced form interest rate equation (14) for the whole period, 1952–70.

It may seem that with the help of the adaptive expectations hypothesis the observed breaks in equations (15) and (16) can be explained by a break in the equation (6). To test a simple proposition that there was a parallel shift in the relation (6) after 1960, I estimated the whole model (6)–(8) after introducing a (0–1) dummy variable in equation (6). The coefficients appeared with right signs and the value of $\beta$ was 0.944. But the improvement in the fit on $\overline{R}^2$ criterion was only

marginal; it was apparent that the introduction of the dummy variable was not adequate to explain the shift in equation (6).

With extrapolative and distributed-lag hypotheses, there is of course the real possibility that the break occurred in both the relations (7) and (8) around 1960. Accordingly, we estimated the model with all the four hypotheses on the formation of expectations for the periods 1952–59 and 1960–70 separately.[2] The estimates were very similar to those obtained by Gibson (1972) and Turnovsky (1970); and support the hypothesis that after 1960 people have become more conscious in forming their price expectations and relating them to nominal interest rates.

## IV. Conclusions

Without being unduly repetitive, I will close by indicating briefly what I believe to be the main conclusions that can be drawn from the material presented.

### *A*

To avoid certain problems associated with the use of a deterministic distributed lag function of past prices as an observable proxy for inflationary expectations, Gibson and Pyle used some survey data on price expectations. To substantiate the speculation that the data may contain errors, I proposed an Unobservable Variable Model which accommodates not only some errors in measurement but also the presence of stochastic error term in the distributed lag function determining the unobservable variable. Evidence was found that the data on short-term expectations contain substantial error in measurement. This explains why both Gibson and Pyle obtained such low values for the coefficient corresponding to the price speculations variable in the Fisher equation.

### *B*

The evidence that observed price expectations data contain errors does not render them unusable. Occasional abrupt changes in price expectations can best be "tracked" by such data. The unobservable variable model, by a judicious use of both the observed price expectations data and the distributed lag proxy, produces many precise and sensible estimates of all the structural parameters.

### *C*

Turnovsky used the same data to test three hypotheses on the formation of price expectations in terms of past prices. In the context of my structural model more precise estimation of individual parameters were possible.

### *D*

Frenkel (1975a,b) has proposed a unified theory on the formation of short-term expectations which is supposed to be consistent with the empirical evidence on the effects of monetary expansion in the short run as well as in the intermediate and long run. It is alleged that Cagan's model of inflationary process does not explain the transitory period between two long-run equilibrium situations. In Frenkel's formulation, short-term expectations are determined by both regressive and adaptive elements. We found that the hypothesis explains the available data better than the other three hypotheses considered by Turnovsky. The short-term expectations alone explain more than 70 percent of the variations in the nominal interest rate. Though some effects of price expectations on real rate can never be overemphasized, I did not pursue that point in this paper.

[2] While estimating the model for the subperiod 1960–70, I also used the 6-month Treasury Bill rates in place of 3-month rates. The results were not found to be significantly different.

### E

My calculations support earlier findings by Gibson and Turnovsky that both the interest rate equation and the expectations formation equation had a structural break around 1960. Yohe and Karnosky found the same in the context of their regressions of nominal interest rates on distributed lag of past prices. However, when I used the adaptive expectations hypothesis, the Chow test of equality of regressions did not support a break in the reduced form interest rate equation.

## REFERENCES


P. D. Cagan, "The Monetary Dynamics of Hyperinflation," in M. Friedman, ed., *Studies in the Quantity Theory of Money*, Chicago 1956.

I. Fisher, *The Theory of Interest*, New York 1930.

J. A. Frenkel, (1975a) "Inflation and the Formation of Expectations," *J. Monetary Econ.*, Oct. 1975, *1*, 403–22.

———, (1975b) "Inflationary Expectations and Some Dynamic Aspects of the Welfare Cost," in M. Parkin and G. Zis, eds., *Inflation in the World Economy*, Manchester 1975.

W. E. Gibson, "Interest Rates and Inflationary Expectations," *Amer. Econ. Rev.*, Dec. 1972, *62*, 854–65.

R. J. Gordon, "Inflation in Recession and Recovery," *Brookings Papers*, Washington 1971, *3*, 105–58.

K. Lahiri, "Estimation of Econometric Models with Unobservable Variables," unpublished doctoral dissertation, Univ. Rochester 1975.

D. H. Pyle, "Observed Price Expectations and Interest Rates," *Rev. Econ. Statist.*, Aug. 1972, *54*, 275–80.

T. J. Sargent, "A Note on the 'Accelerationist' Controversy," *J. Money, Credit, Banking*, Aug. 1971, *3*, 721–25.

S. J. Turnovsky, "Empirical Evidence on the Formation of Price Expectations," *J. Amer. Statist. Assn.*, Dec. 1970, *65*, 1441–54.

——— and M. L. Wachter, "A Test of the 'Expectations Hypothesis' Using Directly Observed Wage and Price Expectations," *Rev. Econ. Statist.*, Feb. 1972, *54*, 47–54.

W. P. Yohe and D. S. Karnosky, "Interest Rates and Price Level Changes," *Fed. Reserve Bank St. Louis Rev.*, Dec. 1969, *51*, 19–36.

# A Model of the Supply of Bilateral Foreign Aid

By LEONARD DUDLEY AND CLAUDE MONTMARQUETTE*

Despite the attention that has been paid to foreign aid in recent years, there has been very little empirical research to explain actual transfers of public funds between donors and recipients. Many authors have indeed discussed the possible motives of donor countries[1] in granting aid, but most empirical work has been directed either toward estimating aid needs or toward assessing the impact of aid on recipient countries. This study approaches the question by considering aid as a good which is consumed indirectly by the residents of the *donor* country. The supply of foreign aid is then explained by the *demand* by the donor country for the impact its aid has on the recipient country. The first step is to develop a model based on the utility function of decision makers of the donor country. This model is then applied to the 1970 aid commitments of fifteen countries belonging to the Development Assistance Committee of the Organization for Economic Cooperation and Development (*OECD*).

* Université de Montréal. Research for this study was financed by grants from the Canada Council (S72-2248) and from the Quebec Ministry of Education. Pierre Bélanger, Gilbert Boileau, and Robert Dépatie undertook a preliminary exploration of the subject in an undergraduate course paper. The computer program for Probit Analysis was supplied by John G. Cragg. We would like to thank M. G. Dagenais for helpful suggestions and Francine Brisson, Assame Diop, Aimé Gogué, and Roger Pruneau for their assistance in conducting the research. Remaining errors are the sole responsibility of the authors.

[1] I. M. D. Little and J. M. Clifford, Raymond Mikesell, George Abbott, Keith Griffin and John Enos, and Janos Horvath have all suggested several possible motives for giving foreign aid: political and economic considerations as well as the simple desire to help the less fortunate.

## I. Motives for the Supply of Foreign Aid

The question of why aid is given is not as easily answered as it might once have been. As recently as 1960, for example, the bilateral aid of the United States, France, and the United Kingdom accounted for over 80 percent of the total flow of bilateral and multilateral aid. One might have been tempted to summarize the reasons for this aid in two terms: cold war and colonialism. Ten years later, however, this share had fallen to about 50 percent of the total, as the aid of the three major donors stagnated while other countries whose political motives were perhaps less overt increased their aid rapidly. Here then is an economic phenomenon to be explained.

But the analysis of foreign aid flows is not simply an academic exercise. The end results of this aid-granting process could scarcely be considered equitable, with a number of moderately prosperous developing countries receiving 20 to 100 times as much aid per capita in 1970 as a number of other extremely poor countries. If the equitable distribution of aid may be taken as an objective, then a necessary starting point is the explanation of the present complex network of bilateral aid flows.

One characteristic of foreign aid flows which has been noted previously is what the *OECD* (p. 178) has referred to as the "small country effect"; that is, the marked tendency for small countries to receive more aid per capita than large countries. The *OECD* estimated that between 1960 and 1967 the eighty-two developing coun-

tries it covered received \$23.7 million per country plus an additional \$1.90 for each resident. In attempting to explain this bias, the *OECD* pointed to the greater openness of small countries and their consequently greater requirements of aid for financing their imports. In short, the bias toward small countries was seen as a demand phenomenon, explained by their higher aid needs.

It must be realized, however, that the demand side is only part of the story, and that actual flows must be explained by the interaction of the aid demand of recipient countries and the aid supply of donor countries. In an empirical context, it is by no means a simple matter to identify separately each of the two sides of the relationship. As we explain below, the supply side is perhaps the easier to approach, since there exist data on foreign aid commitments, which are presumably relatively free of the influence of demand.[2] What we will try to do in this study, then, is to develop a theory of the supply of foreign aid. Our basic assumption is the following: people usually give because they expect to get something in return. What is given in return is invariably something intangible and difficult to quantify; otherwise the transfer would be called an exchange rather than a gift. But in practice very few transfers are unilateral.

At an individual level there may well be exceptions to this generalization. There may indeed exist truly altruistic people who give anonymously and who expect in return neither gratitude from the recipient nor recognition from their fellows. In society as a whole, however, casual empiricism leads us to believe that such individuals are likely to be in a minority. Therefore, any decision to transfer public funds from one society to another will probably be made for motives other than altruism.

What do people expect to get in return when they make public transfers to another society? Various authors have suggested that foreign aid is motivated by political, economic, and charitable considerations.[3] We may perhaps summarize these expectations as follows:

(i) Decision makers in the donor country may expect that the recipient nation will behave more favorably toward their country, lending its support to their national political interests.

(ii) They may expect that the recipient country will confer economic benefits on their country; for example, by buying more of the products they export.

(iii) They may expect some indication that they have had a favorable impact on the residents of the recipient country; perhaps some expression of gratitude, or simply the evidence that the lives of these people have been changed by their charity.

An important point follows: the volume of each of these return flows, whether tangible or intangible, may be supposed to depend in part on the original flow of foreign aid to the recipient. In other words the residents of the donor country may be said to consume their own aid indirectly through the effects it produces in the recipient country. We may, therefore, conclude that the supply of foreign aid will be explained by the donor country's *demand* for foreign aid impact. The next step is to develop a theoretical structure for the supply of foreign aid, suitable for econometric analysis.

## II. Two Models of the Supply of Foreign Aid

In this section we shall consider two versions of a model of the supply of

[2] The comparison to be made is between commitment data and actual disbursement data, which presumably reflect to a much greater extent the interaction of supply and demand.

[3] See fn. 1.

foreign aid. In the first version, there will be no administrative costs. In the second version, there will be positive administrative costs.

### A. *Model 1: No Administrative Costs*

In this first version without administrative costs, we may think of the donor as simply filling out a check and mailing it to the beneficiary. Assume that there is only one good other than foreign aid. Let the decision makers of the donor country have the following utility function:

$$U = f(X, H) \tag{1}$$

where $H$ = consumption of the subjectively measured impact of foreign aid, treated as a private good,[4] and $X$ = total consumption of the other good.

Let $H$ be the sum of the impacts of the donor's aid on $m$ receiving countries

$$H = \sum_{j=1}^{m} H_j = \sum_{j=1}^{m} H(n_j, a_j, y_j) \tag{2}$$

where $H_j$ = subjectively measured impact on beneficiary $j$
$n_j$ = population of beneficiary $j$
$a_j$ = aid per capita received by country $j$
$y_j$ = per capita $GNP$ of country $j$

The impact on country $j$ will be an increasing function of $j$'s population and the per capita aid it receives, and a decreasing function of $j$'s per capita $GNP$:

$$\frac{\partial H}{\partial n_j} > 0, \qquad \frac{\partial H}{\partial a_j} > 0, \qquad \frac{\partial H}{\partial y_j} < 0$$

More specifically, let the impact function take the following form:

$$H_j = n_j^{\alpha}\left(\frac{a_j}{y_j}\right)^{\gamma} \quad 0 \leq \alpha \leq 1, \quad 0 \leq \gamma < 1 \tag{3}$$

The right side of (3) is essentially the relative impact of the aid on the beneficiary's economy weighted by that country's population. The elasticity $\alpha$ is then a measure of the distortion in the donor's perception of the impact of its aid, due to the division of the world's poor into countries of different size (in terms of population). If $\alpha = 1$, the "impact" of aid in each country is weighted by that country's population (no distortion); if $\alpha = 0$, the impacts are simply added without any weighting (complete distortion); if $0 < \alpha < 1$ distortion is only partial. As for the elasticity $\gamma$, it is an indication of the extent of decreasing returns in the creation of impact. Note that if $\gamma = 1$ there are constant returns, and therefore no reason why the donor should not give all its aid to a single poor country.[5]

Finally, the donor country is limited by its budget constraint

$$X + \sum_{j=1}^{m} n_j a_j = Y \tag{4}$$

where $Y = GNP$ of donor.

The maximization of $U$ subject to equations (2), (3), and (4) yields the following first-order condition:

$$\frac{f_j}{f_x} = \frac{y_j^{\gamma} n_j^{1-\alpha}}{\gamma a_j^{\gamma-1}} \tag{5}$$

[4] An alternative formulation which we have explored is to treat aid as a public good. In this case, the utility function (1) becomes that of the individual resident of the donor country, who consumes his share of the private good plus *all* of the public good, foreign aid impact. When incomes are unequal, the median voter decides in effect for everybody. The implication of this approach is that other things being equal, a larger donor should give less aid per capita than a smaller donor since the cost of the public good is shared by a greater number of taxpayers. However, the statistical results—in particular, the sensitivity of the constant term in equation (19) below to the donor's population—did not support the public good hypothesis. For simplicity, therefore, we present the argument in its private good version only.

[5] It will be noted that $\partial^2 H_j/\partial y_j^2$ is nonnegative; that is, the negative effect of an increase in income on aid impact levels off as income grows. The simplest explanation is that the charity component mentioned earlier rises increasingly sharply as income falls toward the subsistence level.

where $f_j$ and $f_x$ are the first-order partial derivatives of $U$ with respect to $H_j$ and $X$. Now the right side of equation (5) may be interpreted as the price of weighted foreign aid impact $H_j$ in terms of other goods foregone. But the weighted impact on recipient $j$ is a perfect substitute for the weighted impact of aid on any other recipient. Therefore, in equilibrium the price of weighted impact in each recipient should be identical:

$$\left(\frac{1}{\gamma}\right) a_j^{1-\gamma} y_j^{\gamma} n_j^{1-\alpha} = k, \qquad j = 1, \ldots, m$$

where $k$ is the marginal rate of substitution between aid impact and the other good. The amount of per capita aid that will be given to country $j$ is then:

$$(6) \qquad a_j = \left[\frac{\gamma k}{y^{\gamma} n_j^{1-\alpha}}\right]^{1/(1-\gamma)}$$

It should be noted that one of the characteristics of this first model is that it implies that a given donor will grant a positive amount of aid to every country.

### B. *Model 2: Positive Administrative Costs*

Consider now a somewhat more realistic version with positive administrative costs. Replace the budget constraint (4) by:

$$(7) \quad X + \left[\sum_{j=1}^{m} a_j n_j - c(a_j n_j)^{\delta}\right] = Y$$

$$0 < \delta < 1$$

where $c$ is a constant.

The second term in parentheses in equation (7) expresses the aid program's administrative costs. With $0<\delta<1$, we are assuming that these costs increase with the amount of aid, but less than proportionally.

Maximizing (1) subject to (2), (3), and (7) now yields:

$$(8) \quad a_j^{1-\gamma} + c\delta n_j^{\delta-1} a_j^{\delta-\gamma} - k\gamma n_j^{\alpha-1} y_j^{-\gamma} = 0$$

To simplify the algebra assume that $\delta$, the elasticity of these costs in relation to the amount of aid, takes on the same value as $\gamma$, the parameter which expresses the decreasing returns in the creation of impact. In this case, the solution of (8) is:

$$(9) \quad a_j = (k\gamma n_j^{\alpha-1} y_j^{-\gamma} - c\gamma n_j^{\gamma-1})^{1/(1-\gamma)}$$

The amount of aid granted will be positive if the expression in parentheses in equation (9) is positive; that is, if

$$(10) \qquad n_j^{\alpha-\gamma} > c y_j^{\gamma}/k$$

If the scale elasticity is small compared to the distortion parameter ($\delta=\gamma<\alpha$), the amount of aid will be zero for populations smaller than

$$n_1 \equiv (c y_j^{\gamma}/k)^{1/(\alpha-\gamma)}$$

and positive for larger populations.

Moreover, the right side of equation (9) reaches a maximum at

$$n_2 \equiv n_1[(1-\gamma)/(1-\alpha)]^{1/(\alpha-\gamma)}$$

The amount of aid per capita will thus be zero for populations smaller than $n_1$, will rise with the beneficiary's population from $n_1$ to $n_2$, and will fall thereafter.

It should be noticed, however, that there are two other possibilities. Should the three parameters be equal ($\delta=\gamma=\alpha$), the amount of aid per capita will be a constant (possibly zero), regardless of the beneficiary's population. But if $\delta=\gamma>\alpha$, the amount of aid given will be a decreasing function of the beneficiary's population, becoming zero for populations equal to or greater than $n_1$.

The results of this section are not necessarily limited to the particular case in which $\delta=\gamma$. Provided that the cost elasticity $\delta$ is less than a critical value, no aid will be granted to countries below a certain population. Should $\delta$ be greater than this critical value no aid will be given to

countries with more than a certain population.

### C. *Some Implications of the Models*

We may summarize the implications of the two models with the help of Table 1 and Figure 1. In effect, there are two decisions which must be taken. First, there is the decision whether or not to grant aid. In model 1, this decision is always taken in the affirmative. In model 2, however, the probability that a given donor will grant any aid at all to country $j$ may depend on the country's income level and population:

$$(11) \qquad Pr\,(a_j > 0) = g(y_j, n_j)$$

The partial derivative of $g$ with respect to $n_j$ may be either positive, negative, or zero as shown in the preceding section. But condition (10) indicates that the partial derivative with respect to the beneficiary's income level is negative:

$$(12) \qquad \partial g/\partial y_j < 0$$

TABLE 1—PROBABILITY OF GRANTING AID AND EFFECT OF RECIPIENT'S POPULATION ON THIS PROBABILITY AND ON AMOUNT GRANTED[a]

| Model | $Pr\ g=(a_j>0)$ | $\partial g/\partial n_j$ | $\partial a_j/\partial n_j \mid a_j>0$ |
|---|---|---|---|
| *Model 1* | 1 | 0 | – |
| *Model 2* | | | |
| (i) $\delta=\gamma<\alpha$ | $0<g<1$ | + | ± |
| (ii) $\delta=\gamma=\alpha$ | $g=0$ or $g=1$ | 0 | 0 |
| (iii) $\delta=\gamma>\alpha$ | $0<g<1$ | – | – |

[a] $a_j$=per capita aid granted to country $j$
$n_j$=population of recipient $j$
$\alpha$=elasticity of aid impact with respect to recipient's population
$\gamma$=elasticity of aid impact with respect to aid as proportion of recipient's *GNP*
$\delta$=elasticity of administrative costs with respect to amount of aid granted

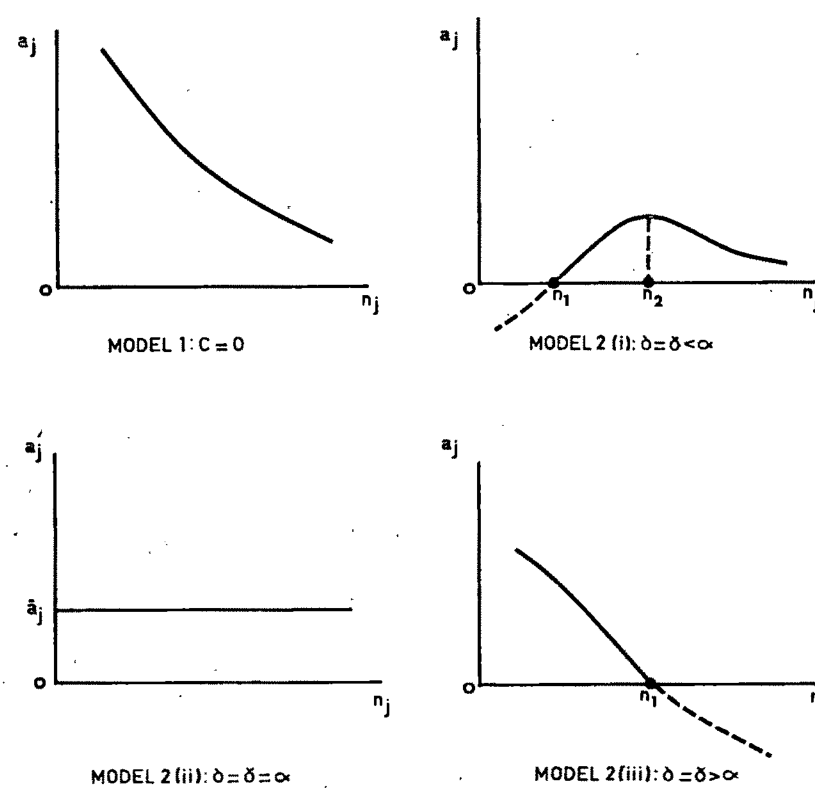



FIGURE 1. EFFECT OF RECIPIENT'S POPULATION ON PER CAPITA AID GRANTED

Given that aid will be granted to country $j$, the question is then how much to give. Model 1 answers this question with equation (6). Model 2 uses a more complicated formula given by equation (9). If $\delta=\gamma=<\alpha$, however, it may be seen that beyond the maximum at $n_2$, equation (9) approaches equation (6). We may therefore use equation (6) to explain the amount of aid granted, if this amount is positive, noting that in the case of model 2 the amount of aid need not necessarily decrease with the beneficiary's population.

### D. *Auxiliary Variables*

There are a number of factors other than the ratio of aid to $GNP$ which might be expected to have an effect on the impact of foreign aid in the recipient country, as perceived by the donor. For example, one factor is the existence of previous political ties between the two countries. One would expect that where these ties exist, a given level of aid would have a greater political impact than where no ties exist. Similarly, one might expect the economic returns to the donor from a given amount of aid to be greater where there are already economic links between the two countries. Finally, there might be a "bandwagon effect" whereby the donor's residents give a higher evaluation to the impact of their aid, the greater the aid the rest of the world gives to the recipient country. This effect would be analogous to Harvey Leibenstein's bandwagon effect.

These changes may be incorporated into the production function for impact, equation (3).

$$H_j = n_j^{\alpha}\left[\frac{a_j}{y_j}\right]^{\gamma}(e_j+1)^{\mu}w_j^{\phi}\prod_{i=1}^{r}C_i^{Z_i} \tag{13}$$

where $Z_i=0$, if there are no political links of type $i$ between donor and receiver, $i=1,\ldots,r$

$Z_i=1$, if there are political links of type $i$, $i=1,\ldots,r$

$e_j=$ previous year's exports of the donor to recipient $j$, per capita of $j$

$w_j=$ aid of the rest of the world to country $j$

In model 2, these auxiliary variables will affect the probability of granting aid. Equation (11) becomes

$$Pr\,(a_j>0) = g(y_j, e_j, w_j, Z_1, \ldots, Z_r) \tag{14}$$

where

$$\partial g/e_j>0,\quad \partial g/\partial w_j>0,\quad \partial g/\partial Z_i>0,\quad i=1,\ldots,r \tag{15}$$

The auxiliary variables will also affect the amount of aid given. Equation (6) must be replaced by

$$a_j = \left[\frac{k\gamma(e_j+1)^{\mu}w_j^{\phi}\prod_{i=1}^{r}C_i^{Z_i}}{y_j^{\gamma}n_j^{1-\alpha}}\right]^{1/(1-\gamma)} \tag{16}$$

In the light of the above discussion, we would expect the following values for the parameters of equation (16):

$$0\le\gamma<1,\quad \mu>0,\quad 0\le\alpha\le1,\quad \phi>0,\quad C_i>1,\quad i=1,\ldots,r \tag{17}$$

All that remains now is to develop a specification convenient for empirical estimation. First there is the decision whether or not to give aid. We replace equation (14) by:

$$A_j = b_0+\overset{-}{b}_1\,ln\,y_j+\overset{?}{b}_2\,ln\,n_j + \overset{+}{b}_3\,ln\,(e_j+1)+\overset{+}{b}_4\,ln\,w_j + \sum_{i=1}^{r}\overset{+*}{b}_iZ_i+\omega_j \tag{18}$$

where $A_j=0$ if no aid is given to country $j$, $A_j=1$ if aid is given to country $j$, and $\omega_j$ is a random disturbance. Expected signs for the coefficients of (18) are derived from conditions (12) and (15).

Turning now to the amount of aid to be given once the recipients have been chosen, we may take the natural logarithm of equation (16) and add an error term:

$$(19)\quad \ln a_j = \beta_0 + \overset{-}{\beta}_1 \ln y_j + \overset{?}{\beta}_2 \ln n_j + \overset{+}{\beta}_3 \ln (e_j + 1) + \overset{+}{\beta}_4 \ln w_j + \sum_{i=1}^{r} \overset{+*}{\beta_i} Z_i + \epsilon_j$$

where

$$\beta_0 = \frac{\ln (k\gamma)}{1-\gamma} \quad \beta_1 = \frac{-\gamma}{1-\gamma} \quad \beta_2 = \frac{-(1-\alpha)}{(1-\gamma)}$$

$$\beta_3 = \frac{\mu}{(1-\gamma)} \quad \beta_4 = \frac{\phi}{(1-\gamma)} \quad \beta_i^* = \frac{\ln C_i}{1-\gamma}$$

$$i = 1, \ldots, r$$

$\epsilon_j$ is an error term with $E(\epsilon_j)=0$, $E(\epsilon_j^2)=\sigma^2$, for all $j$, and the expected signs of the coefficients follow from our hypotheses concerning the values of the structural parameters as presented in condition (17).

The next step is to estimate equations (18) and (19). For (18) we used probit analysis and for (19) ordinary least squares.

## III. The Data

As we have already noted, we decided to use data on (gross) commitments of foreign aid (1970). These commitment data appear to be a better proxy for foreign aid supply than the alternative disbursement data, which are more likely to represent the results of a compromise between the aid demand of recipient countries and the aid supply of donor countries. As a measure of foreign aid, we took the *OECD*'s definition of Official Development Assistance, which includes grants and those loans which have a grant equivalent of 25 percent or more of their nominal value.

A major problem was to decide which countries to include as beneficiaries for a given donor. A first decision was to eliminate colonies and dependencies from our sample, on the grounds that aid to these regions constituted an internal transfer and thus could not be explained by the same model as that used for transfers between countries. This left ninety-three independent developing countries who had received aid from one or more of the fifteen donors in our sample.

As specified in equation (13), dummy variables were used as proxies for political links between donor and recipient countries. However, the political links included varied somewhat from one donor to another. For France, the United Kingdom, and Belgium, we included dummies which specified whether or not the recipient was a former colony of the donor. For Japan and the United States, the dummy indicated whether or not the country was in Asia or in the Western Hemisphere, respectively. In the case of Canada, one dummy corresponded to former British colonial status and the other to former French colonial status. For Belgium, we also tried the dummy corresponding to former French colonial status. Finally for Australia, Denmark, Norway, and Sweden, the dummy corresponded to former British colonial status and for the Netherlands to former Belgian colonial status.

The foreign aid commitment data for 1970 are unpublished figures compiled by the *OECD*. Data on population and Gross National Product per capita in *U.S.* dollars for 1970 were taken from International Monetary Fund and World Bank Group, *Finance and Development*, 10 (March 1973). Export data for 1969 were taken from International Monetary Fund, *Direction of Trade* and are in millions of *U.S.* dollars.

## IV. Statistical Results

Our first step was to use probit analysis to estimate equation (18). The results presented in Table 2 are quite satisfac-

TABLE 2—PROBIT ANALYSIS OF THE MODEL OF THE SUPPLY OF FOREIGN AID[a]

| | Coefficient | | | | | | | | | | | Number of observations | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Donor | $b_0$ | $b_1$ | $b_2$ | $b_3$ | $b_4$ | $b_1^*$[b] | $b_2^*$[b] | $b_3^*$[b] | $b_4^*$[c] | $b_5^*$[c] | $R^2$[d] | Positive | Nil |
| Australia | −2.09 | −1.28[e] | .77[e] | .20[e] | .88[e] | 1.12[e] | | | | | .59 | 21 | 72 |
| | | (.41) | (.23) | (.009) | (.36) | (.45) | | | | | | | |
| Austria | .19 | −.81[e] | .63[e] | .27 | .19 | | | | | | .37 | 17 | 76 |
| | | (.37) | (.17) | (.17) | (.22) | | | | | | | | |
| Belgium | −6.98 | .25 | 1.27[e] | .058 | .54[e] | | 5.96 | 5.72 | | | .69 | 67 | 26 |
| | | (.48) | (.30) | (.25) | (.20) | | (30.5) | (71.2) | | | | | |
| Canada | .44 | −.68[e] | .26 | .02 | .34[e] | 1.58[e] | 1.82[e] | | | | .54 | 56 | 37 |
| | | (.32) | (.15) | (.11) | (.18) | (.40) | (.50) | | | | | | |
| Denmark | −3.82 | −.17 | .49[e] | −.043 | .42[e] | .58 | | | | | .26 | 17 | 76 |
| | | (.33) | (.16) | (.18) | (.24) | (.35) | | | | | | | |
| France | −1.71 | −.63[e] | .62[e] | .68[e] | −.03 | | 1.29[e] | | | | .61 | 37 | 56 |
| | | (.35) | (.17) | (.25) | (.15) | | (.66) | | | | | | |
| Italy | .37 | −.58[e] | .58[e] | .38[e] | .16 | | | | | | .37 | 72 | 21 |
| | | (.29) | (.18) | (.12) | (.18) | | | | | | | | |
| Japan | −4.99 | −.87[e] | 1.03[e] | .32 | .64[e] | | | | | 1.62[e] | .70 | 19 | 74 |
| | | (.40) | (.28) | (.22) | (.34) | | | | | (.53) | | | |
| Netherlands | −4.85 | .024 | .73[e] | −.02 | .32 | | | .14 | | | .38 | 14 | 79 |
| | | (.35) | (.20) | (.25) | (.25) | | | (1.15) | | | | | |
| Norway | −2.27 | −1.17[e] | .62[e] | .19 | .84[e] | .87[e] | | | | | .44 | 16 | 77 |
| | | (.41) | (.18) | (.12) | (.37) | (.42) | | | | | | | |
| Sweden | 2.75 | −1.33[e] | .44[e] | .68[e] | −.009 | .33 | | | | | .35 | 14 | 79 |
| | | (.49) | (.16) | (.25) | (.17) | (.39) | | | | | | | |
| Switzerland | −.12 | −.79[e] | .45[e] | .05 | .59[e] | | | | | | .39 | 55 | 38 |
| | | (.34) | (.14) | (.17) | (.17) | | | | | | | | |
| U.K. | −.03 | −.42 | .22 | .22 | .25 | 1.00[e] | | | | | .20 | 78 | 15 |
| | | (.34) | (.17) | (.19) | (.15) | (.55) | | | | | | | |
| U.S.A. | 11.73 | −2.51[e] | −.66 | .36[e] | .57[e] | | | | 3.12[e] | | .70 | 84 | 9 |
| | | (.79) | (.34) | (.21) | (.26) | | | | 1.66) | | | | |

[a] The model: $A_j = b_0 + b_1 \ln y_j + b_2 \ln n_j + b_3 \ln (c_j+1) + b_4 \ln w_j + \Sigma_{i=1}^{r} b_i^* Z_i + \omega_j$.
[b] Corresponding dummy equal to one if recipient is former colony of U.K. ($b_1^*$), France ($b_2^*$), or Belgium ($b_3^*$); otherwise zero.
[c] Corresponding dummy equal to one if recipient is in Western Hemisphere ($b_4^*$) or Asia ($b_5^*$); otherwise zero.
[d] Pseudo $R$-square.
[e] Indicates coefficient significantly different from zero at the .05 level of significance; one- or two-tail test when appropriate.

tory.[6] With respect to $b_1$, $b_3$, $b_4$, and $b_i^*$, only six of the fifty-five coefficients had their sign contradicting expectations; all of them, however, were not significantly different from zero at a .05 level of significance. Of the forty-nine coefficients with the expected sign, thirty-one were significantly different from zero.

The results for the remaining coefficient $b_2$ are especially interesting since they permit us to distinguish among the two models proposed in Section II. Here, eleven of the fourteen coefficients corresponding to the beneficiary's population were significantly different from zero, all with a positive sign. These results strongly support model 2 with positive administrative costs, and seem to support subcase (i) in which the distortion parameter is relatively high—that is, the degree of distortion is relatively low.

Our next step was to estimate the reduced form represented by equation (19) by ordinary least squares. The results fall into two groups. In the first group are Japan and six small donor countries,[7] each of which had twenty-one or fewer observations. Only a few estimated coefficients were significantly different from zero (results not shown). We would suggest that the problem here was an insufficient number of degrees of freedom.

In the second group are eight larger donors, each of whom gave to at least thirty-seven countries. Here the results,

[6] We have excluded from the results West Germany for which there was only one zero observation and hence very little to explain with probit analysis.

[7] Australia, Austria, Denmark, the Netherlands, Norway, and Sweden.

TABLE 3—REDUCED FORM COEFFICIENTS FOR MODEL OF THE SUPPLY OF FOREIGN AID[a]

| | Donor | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Coefficient | Belgium | Canada | France | Italy | Switzerland | U.K. | U.S.A. | W. Germany |
| $\beta_0$ | −5.86 | −5.67 | .69 | −.88 | −5.12 | −5.05 | −.94 | −3.87 |
| $\beta_1$ | −.55 | −.47 | −.32 | −1.02[d] | −.40 | −.98[d] | −.88[d] | −.78[d] |
| | (.38) | (.47) | (.23) | (.41) | (.63) | (.33) | (.43) | (.36) |
| $\beta_2$ | .003 | .08 | −.41[d] | −.31 | −.46[d] | −.12 | .06 | .15 |
| | (.19) | (.18) | (.11) | (.19) | (.17) | (.14) | (.17) | (.13) |
| $\beta_3$ | .41[d] | .16 | .53[d] | .63[d] | −.16 | .55[d] | .31[d] | .52[d] |
| | (.21) | (.15) | (.17) | (.19) | (.37) | (.19) | (.12) | (.25) |
| $\beta_4$ | .35 | .61[d] | −.27[d] | −.08 | .57[d] | .59[d] | .47[d] | .52[d] |
| | (.24) | (.31) | (.12) | (.25) | (.30) | (.17) | (.27) | (.15) |
| $\beta_1^*$ (U.K.)[b] | | 1.76[d] | | | | 2.38[d] | | |
| | | (.53) | | | | (.48) | | |
| $\beta_2^*$ (France)[b] | .57 | 1.57[d] | .77[d] | | | | | |
| | (.49) | (.69) | (.44) | | | | | |
| $\beta_3^*$ (Belgium)[b] | 3.97[d] | | | | | | | |
| | (.98) | | | | | | | |
| $\beta_4^*$ (W. Hemisphere)[c] | | | | | | | 1.09[d] | |
| | | | | | | | (.56) | |
| Standard error of estimate | 1.49 | 1.70 | .62 | 1.88 | 1.59 | 1.48 | 1.94 | 1.46 |
| Number of observations | 67 | 56 | 37 | 72 | 55 | 78 | 84 | 92 |
| $\bar{R}^2$ | .41 | .176 | .86 | .179 | .226 | .575 | .104 | .166 |

[a] The model: $\ln a_j=\beta_0+\beta_1 \ln y_j+\beta_2 \ln n_j+\beta_3 \ln (e_j+1)+\beta_4 \ln w_j+\sum_{i=1}^{r} \beta_i^* Z_i+\epsilon_j$.
[b] Corresponding dummy equal to one if recipient is a former colony at country in parentheses; otherwise zero.
[c] Corresponding dummy equal to one if recipient is in region in parentheses; otherwise zero.
[d] Indicates coefficient significantly different from zero at the .05 level of significance; one- or two-tail test when appropriate.

presented in Table 3, are much more satisfactory. In the case of $\beta_1$, $\beta_3$, $\beta_4$, and $\beta_i^*$, twenty-one coefficients out of thirty-one were significantly different from zero, all with the correct sign. A single coefficient out of three incorrect sign coefficients was significantly different from zero. Once again the population coefficient—in this case $\beta_2$—is of special interest. It varies in sign and is significantly different from zero (with a negative sign) in only two of the eight cases. These results tend to support model 2, subcase (i) in which administrative costs were positive but increased considerably less rapidly than the amount of aid. They are thus consistent with the probit analysis.

This last set of results was then used to derive point estimates of the structural parameters and estimates of their variances, as shown in Table 4. The quality of these estimates is somewhat more difficult to interpret since, owing to the non-linear functions that relate these parameters and the least squares estimates, the estimates of the variances are only approximations.[8] In every case, however, the estimate of the parameter $\gamma$ indicates decreasing returns to scale in the creation of impact (that is, $\gamma<1$). An important result is that there is no uniform evidence of distortion in the direction of small beneficiary countries. It will be remembered that such distortion in the relative impact of aid was suggested as a possible explanation of the observed bias of aid towards small countries. Only in the case of France does the distortion param-

[8] For details of the procedure, involving the first term of a Taylor's series, see Arthur Goldberger, A. L. Nagar, and H. S. Odeh.

TABLE 4—STRUCTURAL PARAMETERS FOR MODEL OF THE SUPPLY OF FOREIGN AID

| | Donor | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Parameter | Belgium | Canada | France | Italy | Switzerland | U.K. | U.S.A. | W. Germany |
| Returns to scale ($\gamma$) | .37[a,b] | .32[b] | .24[a,b] | .50[a,b] | .28[b] | .49[a,b] | .47[a,b] | .44[a,b] |
| | (.18) | (.22) | (.13) | (.10) | (.32) | (.09) | (.12) | (.12) |
| Distortion ($\alpha$) | 1.00[a] | 1.06[a] | .69[a,b] | .85[a] | .67[a] | .94[a] | 1.03[a] | 1.08[a] |
| | (.12) | (.13) | (.11) | (.10) | (.19) | (.07) | (.10) | (.07) |
| Economic links ($\mu$) | .26[a] | .11 | .40[a] | .31[a] | −.11 | .28[a] | .16[a] | .29[a] |
| | (.10) | (.09) | (.07) | (.06) | (.31) | (.07) | (.05) | (.09) |
| Bandwagon Effect ($\phi$) | .23 | .42 | −.20[a] | −.04 | .41 | .30[a] | .25[a] | .29[a] |
| | (.15) | (.26) | (.10) | (.12) | (.26) | (.10) | (.14) | (.11) |
| Political links U.K.[c] ($C_1$) | | 3.30[a] | | | | 3.34[a] | | |
| | | (1.94) | | | | (1.33) | | |
| Political links France[c] ($C_2$) | 1.45[a] | 2.89[a] | 1.79[a] | | | | | |
| | (.53) | (1.66) | (.76) | | | | | |
| Political links Belgium[c] ($C_3$) | 12.86 | | | | | | | |
| | (14.20) | | | | | | | |
| Political links W. Hemisphere[d] ($C_4$) | | | | | | | 1.78[a] | |
| | | | | | | | (.53) | |

*Note:* Figures in parentheses are approximate standard errors only and should be interpreted with caution.
[a] Indicates coefficient significantly different from zero at .05 level of significance if estimate of standard error is correct; one- or two-tail test when appropriate.
[b] Indicates coefficient significantly different from one at .05 level of significance if estimate of standard error is correct; one- or two-tail test when appropriate.
[c] Corresponding dummy equal to one if recipient is a former colony of this country; otherwise zero.
[d] Corresponding dummy equal to one if recipient is in this region; otherwise zero.

eter $\alpha$ appear to be significantly different from unity.

How, then, does one explain the *OECD* finding of a strongly significant negative correlation between per capita aid and the population of recipient countries? We found that in numerous cases there was a negative correlation significantly different from zero between population on the one hand, and exports, aid of the rest of the world, and the political dummy variables on the other hand. Since the estimated coefficients of these variables in the reduced form presented in Table 3 are in most cases positive and significantly different from zero, it follows from specification analysis[9] that a model of foreign aid that leaves out these other variables would produce a downward-biased estimate for the coefficient of the population variable.

Among the variables which determine the level of the impact function, the economic links represented by lagged exports from donor to recipient (per capita of the latter) proved to be very important. Only for Canada and Switzerland does the parameter $\mu$ appear to be nonsignificant. The results for the bandwagon effect are less clear: in four of the eight cases the coefficient appears significantly different from zero, but in the case of France it is with a negative sign. Finally, the political links between donor and recipient proved to be highly significant.

## V. Conclusions

This paper has presented two versions of a model of the supply of foreign aid. The model attempted to explain two decisions made by the donor country. First

[9] See Henri Theil, pp. 520–602.

was the decision whether or not to grant aid to a given developing country. The model suggested that the probability of granting aid was a decreasing function of the recipient's per capita income, as well as a function of economic, political, and bandwagon considerations. The empirical results supported these hypotheses and also indicated that perhaps because of administrative costs that increase less rapidly than the amount of aid granted to a given country, the probability of granting aid increases with the potential recipient's population.

Secondly, the model attempted to explain the amount of aid granted once the decision to give to a particular country had been taken. The apparent small country bias in the granting of per capita aid which was observed in an earlier study does not seem to be due primarily to distortion on the part of donor countries in favor of recipients with small populations. Rather it would seem to be mainly a result of misspecification of the foreign aid supply model—in particular, the omission of economic, political, and bandwagon effects. Finally, there appear to be strongly decreasing returns to a donor in converting its foreign aid into impact on a given recipient country.

REFERENCES

**G. C. Abbott**, "Economic Aid as a Unilateral Transfer of Resources," *J. Polit. Econ.*, Nov./Dec. 1970, *78*, 1213–27.

**A. S. Goldberger, A. L. Nagar, and H. S. Odeh**, "The Covariance Matrices of Reduced-Form Coefficients and of Forecasts for a Structural Econometric Model," *Econometrica*, Oct. 1961, *29*, 556–73.

**K. B. Griffin and J. L. Enos**, "Foreign Assistance: Objectives and Consequences," *Econ. Develop. Cult. Change*, Apr. 1970, *18*, 313–27.

**J. Horvath**, "Foreign Economic Aid in the International Encyclopedia of the Social Sciences: A Review Article," *J. Econ. Lit.*, June 1971, *9*, 432–41.

**H. Leibenstein**, "Bandwagon, Snob and Veblen Effects in the Theory of Consumers' Demand," *Quart. J. Econ.*, May 1950, *64*, 183–207.

**I. M. D. Little and J. M. Clifford**, *International Aid*, London 1965.

**R. F. Mikesell**, *The Economics of Foreign Aid*, Chicago 1968.

**H. Theil**, *Principles of Econometrics*, Toronto 1971.

**International Monetary Fund**, *Finance and Development*, Mar. 1973, 10.

———, *Direction of Trade*, 1968–72.

**Organization for Economic Cooperation and Development** (*OECD*), *Development Assistance*, 1969 Review, Paris 1969.

# Harberger's Welfare Indicator and Revealed Preference Theory

*By* W. E. DIEWERT*

In a recent article, Arnold Harberger (1971) developed a criterion for indicating whether the utility of a single consumer has increased or decreased from situation one (where prices $p^1 \equiv (p_1^1, p_2^1, \ldots, p_N^1)$ prevailed and quantities $x^1 \equiv (x_1^1, x_2^1, \ldots, x_N^1)$ were consumed) compared to situation two (where prices $p^2 \equiv (p_1^2, p_2^2, \ldots, p_N^2)$ prevailed and quantities $x^2 \equiv x_1^2, x_2^2, \ldots, x_N^2)$ were consumed). According to Harberger's (1971, p. 788) criterion, the consumer's utility or welfare has increased going from situation one to two if the following expression is positive (or welfare has decreased if it is negative):[1]

$$(1) \quad H(p^1, p^2, x^1, x^2) \equiv p^{1T}(x^2 - x^1) + \tfrac{1}{2}(p^2 - p^1)^T(x^2 - x^1)$$

Harberger develops his welfare indicator as an approximation to the true change in welfare using a Taylor series expansion,[2] but for my purposes, the derivation of his welfare indicator is irrelevant—it is the properties of the indicator itself that we will be concerned with.

The interesting thing to note about expression (1) is that it makes use of the same basic information which is used in the revealed preference theory developed by Paul Samuelson (1947, pp. 146–63) and others;[3] i.e., price and quantity information in two periods. The question then naturally arises: Is the Harberger welfare indicator consistent with the results of revealed preference theory?[4] The main purpose of this note is to answer this question.

The following proposition shows that Harberger's welfare measure has the desirable property of being antisymmetric; i.e., if the Harberger measure indicates that welfare has increased going from situation one to two, then it will also indicate that welfare has decreased going from situation two to one.

PROPOSITION 1:

$$H(p^1, p^2, x^1, x^2) \equiv -H(p^2, p^1, x^2, x^1)$$

The proof is by computation.

Given two sets of price quantity data plus the assumption of maximizing behavior,[5] we can construct Table 1 which

* University of British Columbia. This research was partially supported by National Science Foundation Grant GS-3269-A2 at the Institute for Mathematical Studies in the Social Sciences, Stanford University. I would like to acknowledge my intellectual indebtedness to S. N. Afriat, and I thank G. C. Archibald, K. J. Arrow, L. J. Lau, and a referee for their valuable comments.

[1] Note that $x^T$ denotes the transpose of the column vector $x$ so that $x^T p \equiv \sum_{i=1}^N x_i p_i$ is the inner product between the vectors $x$ and $p$.

[2] Several other authors have used Taylor series expansions or quadratic approximations along with the assumption of optimizing behavior to measure changes in utility; see Arthur Bowley, Ragnar Frisch (1936, p. 27; 1938), Harold Hotelling (1938, Sec. II), Abraham Wald, John Hicks (1946, pp. 331–33), T. Kloek, and Henri Theil (1967, pp. 200–12; 1968).

[3] See Hendrik Houthakker (1950) and the earlier work by A. A. Konyus and Herman Wold (1943, 1944).

[4] Samuelson (1964) commenting on an earlier version of the Harberger (1964) methodology raised the question of its consistency with revealed preference theory:

"Simple consumer's surplus is known to be exact only in the empirically bizarre case of parallel indifference curves (where it happens to be superfluous). The ancient theory of its approximate validity remains a modern mess; but the generalized theory of revealed preference tells precisely what inferences about individual and group preferences can be derived from various ($P$, $Q$) observations." [p. 94]

[5] We are now assuming that $x^r$ is a solution to the maximization $\max_x \{f(x) : p^{rT}x \leq p^{rT}x^r, x \geq 0_N\}$ for $r = 1, 2$ where $f$ is a utility function (or production function)

TABLE 1—REVEALED PREFERENCE TABLE

| | $p^{1T}x^2 \geq p^{1T}x^1$ | $p^{1T}x^2 < p^{1T}x^1$ |
|---|---|---|
| $p^{2T}x^1 \geq p^{2T}x^2$ | **Case (i)** Zone of Indeterminancy | **Case (ii)** $x^1$ revealed preferred to $x^2$ |
| $p^{2T}x^1 < p^{2T}x^2$ | **Case (iii)** $x^2$ revealed preferred to $x^1$ | **Case (iv)** Inconsistent preferences have been revealed |

partitions the observed price quantity data into four mutually exclusive cases. The question now is: Does Harberger's welfare indicator convey the same information as the revealed preference table? Rather surprisingly, Proposition 2 below shows that Harberger's welfare indicator is entirely consistent.

PROPOSITION 2: *Harberger's welfare indicator is consistent with the weak axiom of revealed preference in the sense that if the observed price quantity data is such that cases (ii) or (iii) in the revealed preference table occur, then the Harberger indicator will correctly indicate an increase or decrease in welfare, assuming that the data were generated by utility maximizing behavior.*

Proofs of the nontrivial propositions can be found in the Appendix.

Since Proposition 2 is a rather useful result, it may be worthwhile to provide some further discussion. Sydney Afriat (1967) has shown that a slight generalization of the weak axiom of revealed preference may be stated in terms of the two *cross coefficients* $D_{12}$ and $D_{21}$:

$$(2) \qquad D_{12} \equiv (p^{1T}x^2 - p^{1T}x^1)/p^{1T}x^1$$

$$(3) \qquad D_{21} \equiv (p^{2T}x^1 - p^{2T}x^2)/p^{2T}x^2$$

which is continuous from above (so that the maximum exists) and is subject to local nonsatiation, so that solutions to the utility maximization problem will occur on the boundary of the feasible region rather than in the interior.

Afriat's property of *cyclical consistency* in the case of only two price quantity observations can be stated as: if $D_{12} \leq 0$ and $D_{21} \leq 0$, then $D_{12} = D_{21} = 0$. Thus Afriat's property rules out the case of inconsistent preferences, case (iv) in Table 1. Now from (2) and (3), we see that:

$$(4) \quad H(p^1, p^2, x^1, x^2) \equiv \tfrac{1}{2}(p^{1T}x^1)D_{12} - \tfrac{1}{2}(p^{2T}x^2)D_{21}$$

Thus Harberger's welfare indicator is precisely equal to a weighted sum of Afriat's cross coefficients and the proof of Proposition 2 becomes (perhaps) transparently obvious.

To summarize, I have shown that Harberger's welfare indicator will correctly show whether utility has increased or decreased in cases (ii) and (iii); however, if the data reveal inconsistent preferences (i.e., case (iv) occurs), then formula (1) will tell us that welfare has increased if $H(p^1, p^2, x^1, x^2) > 0$ or decreased if $H(p^1, p^2, x^1, x^2) < 0$ when in fact we should not impute any welfare change at all in view of the fact that inconsistent preferences have been revealed.[6]

Let us suppose that we rule out case (iv); that is, we make some side calculations and check whether $D_{12} < 0$ *and* $D_{21} < 0$ in which case we agree not to use formula (1) (or any formula for that matter) in order to measure welfare changes, since inconsistent preferences have been revealed.

However, ruling out case (iv), it appears at first glance that the Harberger welfare

[6] The referee makes the following valid points:

"To make a welfare comparison of two situations characterized by price and consumption vectors, the very first thing to do is consult Table 1. If cases (ii) or (iii) pertain, then an unambiguous welfare comparison can be made without any reference to a special indicator. If case (iv) pertains, then the crucial hypothesis of a single 'rational' consumer is refuted, and analysis in this vein must cease. Only if case (i) pertains, is revealed preference theory ambiguous, and is recourse necessary to a special indicator, based on special assumptions."

indicator has some rather desirable properties: it is consistent with revealed preference theory and it merely resolves the zone of indeterminancy when the data imply that we are in case (i). The following proposition shows that the Harberger indicator does not do a very good job of resolving the zone of indeterminancy.

PROPOSITION 3: *Suppose that we are in the interior of the zone of indeterminancy; i.e.,* $p^{1T}x^2 > p^{1T}x^1$ *and* $p^{2T}x^1 > p^{2T}x^2$. *Then Harberger's welfare indicator can be made positive or negative simply by scaling prices in either period.*[7]

Proposition 3 indicates that the resolution of the zone of indeterminancy generated by the Harberger welfare indicator is extremely unsatisfactory. However, this unsatisfactory scaling property of the Harberger indicator could be eliminated by using *normalized* prices $v^r \equiv p^r/p^{rT}x^r$ for $r=1, 2$ (i.e., prices divided by expenditure in that period) in place of the prices $p^r$. In this case

$$(5)\quad H(v^1, v^2, x^1, x^2) = H(p^1/p^{1T}x^1, p^2/p^{2T}x^2, x^1, x^2) = \tfrac{1}{2}[v^{1T}x^2 - 1] - \tfrac{1}{2}[v^{2T}x^1 - 1] = \tfrac{1}{2}D_{12} - \tfrac{1}{2}D_{21}$$

and the *normalized Harberger welfare indicator* defined by (5) is independent of the scaling of prices.

Now, let us relate the normalized Harberger measure of welfare change to Hicks' (1946, pp. 330–33) measure of consumer surplus which he defines to be one half of the equivalent variation plus one half of the compensating variation. If we translate Hicks' infinitesimal changes into finite differences, we find that Hicks' measure of consumer surplus $\hat{H}$ is defined as:

$$(6)\quad \hat{H}(p^1, p^2, x^1, x^2) \equiv x^{1T}(p^2 - p^1) + \tfrac{1}{2}(x^2 - x^1)^T(p^2 - p^1) = H(x^1, x^2, p^1, p^2)$$

Thus the Hicksian measure of consumer surplus is of the same general form as Harberger's welfare indicator, except that the role of prices and quantities has been interchanged.

The following proposition shows that any result which can be obtained with respect to the normalized Harberger measure of welfare change has an exact counterpart with respect to the normalized Hicksian measure of welfare change.

PROPOSITION 4: *Hicks' normalized welfare indicator is equal to the negative of Harberger's normalized welfare indicator; i.e.,* $\hat{H}(v^1, v^2, x^1, x^2) \equiv -H(v^1, v^2, x^1, x^2)$.

The proof is by computation.

Suppose that the consumer's true utility function (or the producer's true production function in the context of an *output* indicator) can be represented by some function of $N$ variables, $f(x)$ say. Then we will say that a welfare indicator is *exact* if it correctly determines whether welfare has increased, decreased, or remained constant moving from price quantity situation one to two, relative to the true utility function $f(x)$. Note that exactness is an ordinal property. The following propositions show that the normalized Harberger welfare indicator defined by (5) is much more satisfactory than the unnormalized indicator defined by (1).[8]

[7] Harberger (1971) recognizes that normalized prices should perhaps be used in an inflationary setting: "My own preference as to a conventional way of correcting for changes in the absolute price level is to normalize on net national product = national income" (p. 793).

[8] Note that the normalized indicator defined by (5) will still have the desirable properties given in Propositions 1 and 2 but it will no longer have the undesirable property given by Proposition 3.

PROPOSITION 5: *The normalized Harberger welfare indicator defined by (5) is exact if the true utility function $f(x)$ is a once differentiable monotone transformation of the homogeneous quadratic utility function*

$$x^T Ax \equiv \sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij}x_ix_j$$

*where $a_{ij}=a_{ji}$; i.e., if $f(x)=\hat{g}(x^TAx)$ where $\hat{g}'(z)\equiv d\hat{g}(z)/dz>0$.*

Assume that the consumer maximizes $f(x)$ subject to the budget constraint $v^{*T}x=1$ in the normalized prices $v^*$. Then the first-order necessary conditions for this utility maximization problem yield *Wold's Identity* (1944, pp. 69–71):[9]

$$(7) \qquad v^* = \nabla f(x^*)/x^{*T}\nabla f(x^*)$$

where $x^*$ is a solution to $\max_x \{f(x): v^{*T}x=1,\ x\geq 0_N\}$ and $\nabla f(x^*)\equiv[\partial f(x^*)/\partial x_1, \partial f(x^*)/\partial x_2, \ldots, \partial f(x^*)/\partial x_N]^T$ is the gradient vector of $f$ evaluated at $x=x^*$. We may regard the vector of functions $v(x)\equiv\nabla f(x)/x^T\nabla f(x)$ as the consumer's system of inverse demand functions; that is, $v(x)$ represents the vector of normalized prices which is consistent with $x$ being a solution to the utility maximization problem $\max_z \{f(z): v^T(x)z=1,\ z\geq 0_N\}$.[10] We use Wold's Identity below.

Using the duality between direct and indirect utility functions[11] it is possible to state an analogous version of Proposition 5 in terms of the indirect utility function. Given the direct utility function $f(x)$, one defines the indirect utility function $g$ as follows:

$$(8) \quad g(v) \equiv \max_x \{f(x): v^Tx \leq 1,\ x\geq 0_N\}$$

On the other hand, given an indirect utility function satisfying certain regularity properties, one can define the corresponding direct utility function by:

$$(9) \quad f(x) \equiv \min_v \{g(v): v^Tx \leq 1,\ v\geq 0_N\}$$

The first-order necessary conditions for the minimization problem yield *Roy's Identity*, p. 222:

$$(10) \qquad x^* = \frac{\nabla g(v^*)}{v^{*T}\nabla g(v^*)} \equiv x(v^*)$$

Thus the consumer's system of utility maximizing demand functions $x(v)$ can be generated by differentiating the indirect utility function as in (10) above. We make use of Roy's Identity in the following proposition:

PROPOSITION 6: *The normalized Harberger welfare indicator defined by (5) is exact if the true indirect utility function $g(v)$ is a once differentiable monotone transform of the homogeneous quadratic indirect*[12] *utility function*

$$v^TBv \equiv \sum_{i=1}^{N}\sum_{j=1}^{N} b_{ij}v_iv_j$$

*where $b_{ij}=b_{ji}$; i.e., if $g(v)=\hat{h}(v^TBv)$ where $\hat{h}'<0$.*

[9] Actually, this identity was obtained by Hotelling (1935) and by Konyus and S. S. Byushgens, p. 155, in the case of a linear homogeneous $f$. They also derived a version of Roy's Identity in the linear homogeneous case.

[10] $x\geq 0_N$ means each component of the $N$-dimensional vector is nonnegative; $x>0_N$ means that $x\geq 0_N$ and at least one component is positive; $x\gg 0_N$ means each component of $x$ is positive.

[11] See for example Konyus and Byushgens, René Roy, Houthakker (1952), Lawrence Lau (1969), Ronald Shephard (pp. 301–05), Giora Hanoch, Afriat (1972a), and the author (1974b). Direct utility functions are assumed to be continuous, nondecreasing, and quasi concave; the corresponding indirect utility functions will be continuous, nonincreasing, and quasi convex.

[12] Note that once differentiability of the direct utility function does not necessarily imply that the indirect utility function is differentiable and vice versa. For example, suppose that the indirect utility function is given by $g(v)=1/(v^Tbb^Tv)^{1/2}$ where $b$ is a vector of positive constants. The corresponding direct utility function is a Leontief utility function where the vector $b$ represents the unit utility input-output coefficients. The Leontief utility function is not differentiable.

It is possible to show that the normalized Harberger welfare indicator bears a close relationship to Irving Fisher's ideal index number, pp. 241–42. Let us take the direct utility function $f$ to be a homogeneous quadratic which is normalized so that it is homogeneous of degree one; i.e., define[13]

$$(11) \qquad f(x) = (x^T A x)^{1/2}$$

If the matrix $A$ in (11) is nonsingular, then for positive, normalized price vectors $v$ such that $A^{-1}v > 0_N$, we can show using definition (8) that

$$(12) \qquad g(v) = (v^T A^{-1} v)^{-1/2}$$

In any case, substituting (11) into (7), we find that:

$$(13) \quad 2H(v^1, v^2, x^1, x^2) = x^{1T} A x^2 [\{1/[f(x^1)]^2\} - \{1/[f(x^2)]^2\}]$$

an expression which we will compare with an analogous expression for the *Fisher welfare indicator* defined as:

$$(14) \quad F(v^1, v^2, x^1, x^2) \equiv (v^{1T}x^2/v^{2T}x^1)^{1/2} = ([p^{1T}x^2 p^{2T}x^2]/[p^{1T}x^1 p^{2T}x^1])^{1/2}$$

The following proposition indicates that the Fisher welfare indicator has a nicer cardinal interpretation than the Harberger welfare indicator.

PROPOSITION 7 (Konyus and Byushgens (1926, pp. 167–72)): *If the true utility function can be represented by a direct utility function of the form (11),*[14] *then* $F(v^1, v^2, x^1, x^2) \equiv f(x^2)/f(x^1)$.

The above rather amazing proposition has been noted by Samuelson (1947, p. 155) and by Afriat (1972b, p. 45). The proposition enables us to *cardinally* rank correctly (using a chain index subject to a base period normalization), not only two price-quantity situations, but an arbitrary set of price-quantity situations, provided that the quantities were generated by a single consumer (or producer) maximizing a utility (or production) function of the form (11) subject to an expenditure constraint. Note that the normalized Harberger indicator cannot *cardinally* correctly rank an entire set of data even when $f$ is defined by (11) due to the fact that the magnitude of $x^{1T}Ax^2$ which occurs in (13) is generally unknown, although the normalized Harberger indicator can be used to correctly rank *ordinally* an entire set of data under the stated assumption.

Although the cardinal notion makes good sense in the context of producer theory, "ordinalists" may object to the cardinal concept in the context of consumer theory. Thus below, we prove a version of Proposition 7 that rests only on ordinal preferences. In order to do this, we must define S. Malmquist's concept of a quantity index.

First, given a utility function $f$, define $f$'s *distance function* as $D[u; x] \equiv \max_\lambda \{\lambda : f(x/\lambda) \geq u\}$. The distance function tells us by what proportion one has to deflate the given consumption vector $x$ in order to obtain a point on the utility surface indexed by $u$. In general, $f$ can be completely characterized by $D$ (see Shephard, Hanoch, and Daniel McFadden). In par-

[13] Actually $f(x) = (x^T Ax)^{1/2}$ will not generally be a nondecreasing and quasi-concave function for all $x \gg 0_N$; thus we restrict the domain of definition of $f$ to $x$ such that $x \gg 0_N$, $Ax > 0_N$, and $x$ such that if $y^T Ax = 0$, then $y^T Ay \leq 0$. This will generally imply that $A$ is a matrix with $N-1$ nonpositive eigenvalues and one positive eigenvalue with a positive eigen-vector. This functional form has been studied by Afriat (1972b, p. 45) as a utility function and by the author (1974a) as a factor requirements function (where the quasi-concavity requirement was replaced by a quasi-convexity requirement).

[14] Or by an indirect utility function of the form $g(v) = 1/(v^T Bv)^{1/2}$ where $B$ is a symmetric matrix satisfying certain regularity conditions. The proof for the indirect case is analogous.

ticular, $D[u; x]$ is linearly homogeneous, nondecreasing, and concave in the vector of variables $x$ and nonincreasing in $u$.

Now define the *Malmquist quantity index* as $Q_M(u; x^1, x^2) \equiv D[u; x^2]/D[u; x^1]$. Note that the index depends on $x^1$ (the base period quantities), $x^2$ (the current period quantities), and on the base indifference surface (indexed by $u$) onto which the points $x^1$ and $x^2$ are deflated. The following proposition is free of cardinalist taint.

PROPOSITION 8: *Let the consumer's preferences be represented by the utility function $g[f(x)]$ where $f(x) \equiv (x^TAx)^{1/2}$ satisfies the regularity conditions of footnote 13 and $g'(\lambda) > 0$ for $\lambda > 0$ so that $g$ is a monotonically increasing function of one variable. Suppose the quantity vector $x^1$ is a solution to the utility maximization problem $max_x\{g[f(x)]: p^{1T}x = p^{1T}x^1\}$ while $x^2$ is a solution to $max_x\{g[f(x)]: p^{2T}x = p^{2T}x^2\}$ and $u^1 \equiv g[f(x^1)]$, $u^2 \equiv g[f(x^2)]$. Then the Fisher welfare $F(v^1, v^2, x^1, x^2) = Q_M(u; x^1, x^2)$ (the Malmquist quantity index) for every $u$ (i.e., for any reference indifference surface).*

Propositions 7 and 8 suggest that the Fisher welfare indicator may be interpreted as a quantity index providing that the underlying aggregate $y$ has the function for $y = (x^TAx)^{1/2}$ or is a monotonic transformation of this functional form, irrespective of whether we are in the producer or consumer context. The functional form defined by (11) is "flexible" in the sense that it can provide a second-order approximation to an arbitrary twice differentiable, linear homogeneous function.[15] Thus Fisher's "ideal" index does in fact appear to be reasonably ideal and should be more widely used than it has been.[16]

Returning now to the problem of the determination of the change in welfare when there are only two price-quantity situations, the above results show that the normalized Harberger indicator (5) and the Fisher-Byushgens-Konyus indicator (14) are exact if the underlying true utility function is a monotone transform of a homogeneous quadratic direct or indirect utility function and thus has the property of homotheticity.[17] This property in turn implies that all income elasticities are unitary,[18] which is not consistent with Engel's Law. However, it turns out that the normalized Harberger welfare indicator (5) and the Fisher welfare indicator (14) are exact for functional forms which generate non-linear Engel curves.

We term the Fisher index *exact* for a set of preferences if when $x^1$ is indifferent to $x^2$, $F(v^1, v^2, x^1, x^2) = 1$ and when $x^2$ is preferred to $x^1$, $F(v^1, v^2, x^1, x^2) > 1$. It is clear from (5) and (14) that the class of preferences which is exact for the normalized Harberger welfare indicator is also exact for the Fisher indicator. A partial characterization of the class of preferences for which the Fisher indicator is exact has been provided by Byushgens. He has shown that the indifference surface generated by a differentiable $f$ which has the property: $\{f(x^1) = f(x^2)$ implies $x^{2T}\nabla f(x^1)/x^{1T}\nabla f(x^1) = x^{1T}\nabla f(x^2)/x^{2T}\nabla f(x^2)\}$ is of the form $\{x: x^TAx = 1\}$ for some symmetric matrix $A$. The following extension of Byushgens' result shows that the "quadratic" indifference surfaces can shift in a nonhomothetic manner as the utility level changes.[19]

[15] See the author (1974a) for a proof. Note that the hypothesis of linear homogeneity is essential for this proposition.

[16] Note that the Fisher-Byushgens-Konyus welfare indicator defined by (14) also has the desirable properties given in Propositions 1 and 2; in particular, the indicator is consistent with cases (ii) and (iii) of our revealed preference table.

[17] See Shephard, p. 30, for a definition.

[18] See, for example, Lau (1969, p. 393).

[19] Note that the normalized Harberger indicator is *not* exact for the quadratic utility function $a^Tx + x^TAx$ where $a$ is a nonzero vector and $A$ is a nonzero matrix.

PROPOSITION 9: *Suppose that the preferences of a consumer can be represented by the implicit quadratic distance function,*

$$D[u; x] \equiv \left( \sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij}(u)x_i x_j \right)^{1/2} \equiv (x^T A(u)x)^{1/2}$$

*where* $a_{ij}(u) = a_{ji}(u)$ *and* $(u, x)$ *belongs to a set* $S \equiv \{(u, x): (x^T A(u)x)^{1/2}$ *is a nondecreasing, positive, and concave function of* $x$ *and a decreasing function of* $u\}$. *Let* $x^r$ *be a solution to the utility maximization problem* $\max_x\{f(x): p^{rT}x = p^{rT}x^r, x \geq 0_N\} \equiv u^r$ *and let* $(u^r, x^r) \in S$ *for* $r = 1, 2$ *where* $f$ *is the utility function generated by the implicit quadratic distance function.*[20] *If* $x^2$ *is sufficiently close to* $x^1$ *so that* $x^{2T}\nabla_x D[u; x^1]$ *is a decreasing function of* $u$ *for* $u$*'s in the interval between* $u^1$ *and* $u^2$, *then the Fisher and normalized Harberger indicators are exact.*

Since $D[u; x]$ is linearly homogeneous in $x$, Euler's Theorem implies that $x^{2T}\nabla_x D[u; x^1] = D[u; x^1]$ when $x^1 = x^2$. Since $D[u; x^1]$ is decreasing in $u$ by assumption, it follows that $x^{2T}\nabla_x D[u; x^1]$ will be decreasing in $u$ for $x^2$ close enough to $x^1$. In particular, if preferences are homothetic so that $A(u) \equiv A/g^{-1}(u)$ where $g^{-1}$ is an increasing function of one variable and $A$ is a symmetric matrix satisfying the conditions outlined in footnote 13, then $x^{2T}\nabla_x D[u; x^1]$ will be decreasing in $u$ for every $x^1$ and $x^2$ in the domain of definition defined in footnote 13. In any case, the hypotheses of Proposition 9 do not appear to be particularly restrictive.

It can be shown that there is a dual analogue to Proposition 9; i.e., the normalized Harberger and Fisher indicators are also exact for preferences which can be represented by the *implicit quadratic cost function* $C[u; v] \equiv (v^T B(u)v)^{1/2}$ (for $v^1$ close enough to $v^2$ so that $v^{2T}\nabla_v C[u; v^1]$ is an increasing function of $u$), where $B(u)$ is a symmetric matrix of functions which satisfies the regularity conditions of footnote 13 for each $u$, and in addition $(v^T B(u)v)^{1/2}$ is an increasing function of $u$.

Thus the normalized Harberger and the Fisher welfare indicators are exact for a certain class of nonhomothetic utility functions, and this class of functions appears to be broad enough to provide a second-order differential[21] approximation to an arbitrary twice differentiable direct utility function.

The contents of this paper can be summarized as follows: (i) Harberger's welfare indicator (1) does have some desirable properties (recall Propositions 1 and 2); (ii) in view of Proposition 3, it is better to use the normalized Harberger welfare indicator (5) (or the equivalent Fisher welfare indicator (14) since both indicators give the correct ordinal answer for the same class of preferences); (iii) the normalized Harberger indicator will give the correct answer for a surprisingly broad class of functional forms for utility functions; (iv) Irving Fisher's ideal index number can be given a rather strong economic justification; and (v) Hicks' measure of consumer surplus in finite difference form using normalized prices is equal to the negative of the (normalized) Harberger welfare change.

Although the above results do justify the use of a consumer surplus concept in empirical applications, they do not justify the addition of welfare changes across consumers without some further weighting.[22]

[20] For formal duality theorems see Hanoch, McFadden, and Shephard. The utility function $f$ which corresponds to a distance function $D$ is defined by $f(x) \equiv \max_u \{u: D[u; x] \geq 1\}$.

[21] That is, the approximating function can attain the first- and second-order partial derivatives of the function to be approximated at a given point. This terminology follows Lau (1974).

[22] Note that the consumer surplus technique cannot in general be used to *predict* what the change in welfare would be before the consumer makes his choices. In order to predict, it is necessary to assume a functional

## Appendix

*Proof of Proposition 2:*

Consider case (iii) where $x^2$ is revealed preferred to $x^1$. (The proof of case (ii) is entirely analogous and case (iv) does not occur under the assumption of utility maximizing behavior):

$$\begin{aligned} H(p^1, p^2, x^1, x^2) &= p^{1T}(x^2 - x^1) + \tfrac{1}{2}(p^2 - p^1)^T(x^2 - x^1) \\ &= \tfrac{1}{2}[p^{1T}x^2 - p^{1T}x^1] - \tfrac{1}{2}[p^{2T}x^1 - p^{2T}x^2] \\ &= \tfrac{1}{2}[\text{a nonnegative number}] \\ &\quad - \tfrac{1}{2}[\text{a negative number}] \\ &> 0 \end{aligned}$$

*Proof of Proposition 3:*

Let us scale prices in period two according to the positive number $\lambda$. Then

$$\begin{aligned} H(p^1, \lambda p^2, x^1, x^2) &= \tfrac{1}{2}[p^{1T}x^2 - p^{1T}x^1] \\ &\quad - \tfrac{1}{2}\lambda[p^{2T}x^1 - p^{2T}x^2] \\ &= \tfrac{1}{2}[\text{positive number}] \\ &\quad - \tfrac{1}{2}\lambda[\text{positive number}] \end{aligned}$$

Thus if we choose the scaling factor $\lambda$ to be small, $H>0$ and if we choose $\lambda$ to be large, then $H<0$.

*Proof of Proposition 5:*

$2H(v^1, v^2, x^1, x^2) \equiv [v^{1T}x^2-1]-[v^{2T}x^1-1]=[v^T(x^1)x^2-v^T(x^2)x^1] = \{(x^{2T}\nabla f(x^1))/(x^{1T}\nabla f(x^1))\} - \{(x^{1T}\nabla f(x^2)/(x^{2T}\nabla f(x^2))\} = \{(x^{2T}\hat{g}'(x^{1T}Ax^1) \cdot 2Ax^1)/(x^{1T}\hat{g}'(x^{1T}Ax^1)2Ax^1)\} - \{(x^{1T}\hat{g}'(x^{2T}Ax^2) \cdot 2Ax^2)/(x^{2T}\hat{g}'(x^{2T}Ax^2)2Ax^2)\} = x^{2T}Ax^1\{(1/x^{1T}Ax^1)-(1/x^{2T}Ax^2)\}$ using $A=A^T$. Since $\nabla f(x)>0_N$ and $\hat{g}'>0$, we may restrict our attention to $x\gg 0_N$ such that $Ax>0_N$. In this case, the term $x^{2T}Ax^1$ which occurs above is positive. Thus $H(v^1, v^2, x^1, x^2)>0$ is equivalent to $[1/x^{1T}Ax^1-1/x^{2T}Ax^2]>0$ which is equivalent to (since $x^{rT}Ax^r>0$): $x^{2T}Ax^2> x^{1T}Ax^1$. But since $\hat{g}'>0$, this is equivalent to $\hat{g}(x^{2T}Ax^2)=f(x^2)>f(x^1)=\hat{g}(x^{1T}Ax^1)$. Thus the normalized Harberger welfare indicator correctly indicates that utility has increased moving from the first price-quantity situation to the second. The analysis for the other cases is similar.

form for the direct or indirect utility function and use Wold's Identity (7) or Roy's Identity (10) plus standard econometric techniques in order to estimate the unknown parameters of the utility function. See Laurits Christensen, Dale Jorgenson, and Lau for an empirical application, and the author (1974b) for a review of the literature.

*Proof of Proposition 6:*

$2H(v^1, v^2, x^1, x^2) \equiv [v^{1T}x^2-1]-[v^{2T}x^1-1] = v^{1T}x(v^2)-v^{2T}x(v^1) = [v^{1T}\nabla g(v^2)/v^{2T}\nabla g(v^2)] - [v^{2T}\nabla g(v^1)/v^{1T}\nabla g(v^1)] = v^{1T}Bv^2[(1/v^{2T}Bv^2)-(1/v^{1T}Bv^1)]$ using $g(v)\equiv\hat{h}(v^TBv)$. At this stage, it is necessary to make use of the assumption that $g(v)$ is an indirect utility function, and thus it must be nonincreasing in its arguments; i.e., we assume $\nabla g(v)<0_N$. Since $\hat{h}'<0$, we restrict attention to $v\gg 0_N$ such that $Bv>0_N$ to ensure that $\nabla g(v)<0_N$. Thus the term $v^{1T}Bv^2$ which occurs above is positive and we find that $H(v^1, v^2, x^1, x^2)>0$ is equivalent to $[1/v^{2T}Bv^2-1/v^{1T}Bv^1]>0$ which is equivalent to $v^{1T}Bv^1>v^{2T}Bv^2$ and since $\hat{h}'<0$, this is equivalent to $\hat{h}(v^{2T}Bv^2)=g(v^2)>g(v^1)=\hat{h}(v^{1T}Bv^1)$. Thus the normalized Harberger welfare indicator correctly indicates that utility has increased moving from the first price-quantity situation to the second. The analysis for the other cases is similar.

*Proof of Propositon 7:*

$F(v^1, v^2, x^1,x^2)\equiv(v^{1T}x^2/v^{2T}x^1)^{1/2}=([x^{2T}\nabla f(x^1)/x^{1T}\nabla f(x^1)]/[x^{1T}\nabla f(x^2)/x^{2T}\nabla f(x^2)])^{1/2}$ using (7) $=([x^{2T}Ax^1/\{f(x^1)\}^2]/[x^{1T}Ax^2/\{f(x^2)\}^2])^{1/2}$ using $(11)=f(x^2)/f(x^1)$.

*Proof of Proposition 8:*

If we solve $g[f(x/\lambda)]=u$ for $\lambda$, we obtain the distance function, $\lambda=D[u;x]=(x^TAx)^{1/2}/g^{-1}(u)$ where $g^{-1}[g(u)]\equiv u$.

The first-order conditions for the two utility maximization problems after elimination of the Lagrange multipliers (Wold's Identity) yield the relations $v^1\equiv p^1/p^{1T}x^1=Ax^1/x^{1T}Ax^1$ and $v^2\equiv p^2/p^{2T}x^2=Ax^2/x^{2T}Ax^2$. Use these relations to eliminate $v^1$ and $v^2$ from the Fisher welfare indicator and obtain $F(v^1, v^2, x^1, x^2)\equiv(v^{1T}x^2/v^{2T}x^1)^{1/2}=[(x^{1T}Ax^2/x^{1T}Ax^1)/(x^{2T}Ax^1/x^{2T}Ax^2)]^{1/2}=(x^{2T}Ax^2)^{1/2}/(x^{1T}Ax^1)^{1/2}=[(x^{2T}Ax^2)^{1/2}/g^{-1}(u)]^{1/2}/[(x^{1T}Ax^1)^{1/2}/g^{-1}(u)]^{1/2}=D[u;x^2]/D[u;x^1]\equiv Q_M(u; x^1, x^2)$ for any $u$.

*Proof of Proposition 9:*

It is first necessary to express the partial derivatives of $D$ with respect to the components of $x$, $\nabla_x D[u^r; x^r]$; $r=1, 2$, in terms of the derivatives of $f(x)$. We have $D[u^r; x^r]=1$ for $r=1, 2$ since each $x^r$ is on the utility surface indexed by $u^r$. To determine how the distance $D[u^1; x^1]$ changes as the components of $x^1$ change, apply the implicit function theorem to the equation $f(x^1/\lambda)=u^1$ (where $\lambda=1$ initially). We find that $\partial\lambda/\partial x_j \equiv \partial D[u^1; x^1]/\partial x_j = (\partial f(x^1)/\partial x_j)/x^{1T}\nabla f(x^1)$ for $j=1, 2, \ldots, N$. Similarly, $\nabla_x D[u^2; x^2] = \nabla f(x^2)/x^{2T}\nabla f(x^2)$. Also, by Wold's Identity, $v^1 \equiv p^1/p^{1T}x^1 = \nabla f(x^1)/x^{1T}\nabla f(x^1)$ and $v^2 = \nabla f(x^2)/x^{2T}\nabla f(x^2)$. Upon noting that the right-hand sides of the last two sets of equations are equal to the right-hand sides of the earlier set, we obtain

$$\text{(A1)} \quad v^1 \equiv p^1/p^{1T}x^1 = \nabla_x D[u^1; x^1] \quad \text{and} \quad v^2 \equiv p^2/p^{2T}x^2 = \nabla_x D[u^2; x^2]$$

Now substitute equation (A1) into the Fisher indicator and we obtain $F(v^1, v^2, x^1, x^2) \equiv (v^{1T}x^2/v^{2T}x^1)^{1/2} = (x^{2T}\nabla_x D[u^1; x^1]/x^{1T}\nabla_x D[u^2: x^2])^{1/2} = (x^{2T}\nabla_x D[u^1; x^1]/[x^{1T}A(u^2)x^2/(x^{2T}A(u^2)x^2)^{1/2}])^{1/2} = (x^{2T}\nabla_x D[u^1; x^1]/[x^{2T}A(u^2)x^1])^{1/2}$ (since $D[u^2; x^2] = (x^{2T}A(u^2)x^2)^{1/2} = 1) = (x^{2T}\nabla_x D[u^1; x^1]/\{[x^{2T}A(u^2)x^1x^{1T}A(u^2)x^1]/x^{1T}A(u^2)x^1\})^{1/2} = (x^{2T}\nabla_x D[u^1; x^1]/x^{2T}\nabla_x D[u^2; x^1])^{1/2}/D[u^2; x^1]$. If $u^1 < u^2$, then $x^{2T}\nabla_x D[u^1; x^1] > x^{2T}\nabla_x D[u^2; x^1]$ since $x^{2T}\nabla_x D[u; x^1]$ is a decreasing function of $u$. Also, since $D[u, x^1]$ is decreasing in $u$, $1 = 1/D[u^1; x^1] < 1/D[u^2; x^1]$. Thus $F(v^1, v^2, x^1, x^2) > 1$. Similarly if $u^1 = u^2$, $F(v^1, v^2, x^1, x^2) = 1$ and if $u^1 > u^2$, $F(v^1, v^2, x^1, x^2) < 1$, which proves that the Fisher indicator is exact.

## REFERENCES


S. N. Afriat, "The Construction of Utility Functions from Expenditure Data," *Int. Econ. Rev.*, Feb. 1967, *8*, 67–77.

———, (1972a) "The Case of the Vanishing Slutsky Matrix," *J. Econ. Theory*, Oct. 1972, *5*, 208–23.

———, (1972b) "The Theory of International Comparisons of Real Income and Prices," in D. J. Daly, ed., *International Comparisons of Prices and Outputs*, New York 1972.

A. L. Bowley, "Notes on Index Numbers," *Econ. J.*, June 1928, *38*, 216–37.

S. S. Buscheguennce, "Sur une classe des hypersurfaces: à propos de 'l'index idéal' de M. Irv. Fischer," *Recueil Mathematique* (Moscow), 1925, *32*, 625–31. (Russian title: Byushgens, S. S., "Ob odnom klasse giperpoverkhnostey: po povodu 'idealnovo indeksa' Irv. Fischer'a pokupatelnoi sili deneg," *Matematicheskii Sbornik*, 1925, *32*, 625–631.)

L. R. Christensen, D. W. Jorgenson, and L. J. Lau, "Transcendental Logarithmic Utility Functions," *Amer. Econ. Rev.*, June 1975, *65*, 367–83.

W. E. Diewert, (1974a) "Functional Forms for Revenue and Factor Requirements Functions," *Int. Econ. Rev.*, Feb. 1974, *15*, 119–30.

———, (1974b) "Applications of Duality Theory," in M. Intriligator and D. Kendrick, eds., *Frontiers of Quantitative Economics*, Vol. II, Amsterdam 1974.

I. Fisher, *The Making of Index Numbers*, Boston 1922.

R. Frisch, "Annual Survey of General Economic Theory: The Problem of Index Numbers," *Econometrica*, Jan. 1936, *4*, 1–39.

———, "The Double Expenditure Method," *Econometrica*, Jan. 1938, *6*, 85–90.

G. Hanoch, "Generation of New Production Functions through Duality," Harvard Inst. Econ. Res., 1970.

A. C. Harberger, "The Measurement of Waste," *Amer. Econ. Rev. Proc.*, May 1964, *54*, 58–76.

———, "Three Basic Postulates for Applied Welfare Economics: An Interpretive Essay," *J. Econ. Lit.*, Sept. 1971, *9*, 785–97.

J. R. Hicks, *Value and Capital*, 2d ed., Oxford 1946.

H. Hotelling, "Demand Functions with Limited Budgets," *Econometrica*, Jan. 1935, *3*, 66–78.

———, "The General Welfare in Relation to Problems of Taxation and of Railway and Utility Rates," *Econometrica*, Mar. 1938, *6*, 242–69.

H. S. Houthakker, "Revealed Preference and the Utility Function," *Economica*, May 1950, *17*, 159–74.

———, "Compensated Changes in Quantities and Qualities Consumed," *Rev. Econ. Stud.*, Oct. 1952, *19*, 155–64.

T. **Kloek,** *Indexcijfers: enige methodologische aspecten,* The Hague 1966.

A. A. **Konyus,** "Problema istinovo indeksa stoimosti zhizni," *Ekonomicheskii Byulleten Konyunkturnovo Instituta,* 1924, *3,* 64–71; English translation: "The Problem of the True Index of the Cost of Living," *Econometrica,* Jan. 1939, *7,* 10–29.

——— **and S. S. Byushgens,** "K probleme pokupatelnoi cili deneg," *Voprosi Konyunkturi,* 1926, *2,* 151–72. English title: "On the Problem of the Purchasing Power of Money," *The Problems of Economic Conditions,* Econ. Bull. Conjuncture Inst. suppl., 1926, *2,* 151–71.

L. J. **Lau,** "Duality and the Structure of Utility Functions," *J. Econ. Theory,* Dec. 1969, *1,* 374–96.

———, "Applications of Duality Theory: A Comment," in M. Intriligator and D. Kendrick, eds., *Frontiers of Quantitative Economics,* Vol. II, Amsterdam 1974.

D. **McFadden,** "Cost, Revenue and Profit Functions," Univ. California, Berkeley, 1970.

S. **Malmquist,** "Index Numbers and Indifference Surfaces," *Trabajos de Estadistica,* 1953, *4,* 209–42.

R. **Roy,** "La distribution du revenue entre les divers biens," *Econometrica,* Mar. 1947, *15,* 205–25.

P. A. **Samuelson,** *Foundations of Economic Analysis,* Cambridge, Mass. 1947.

———, Discussion of "The Measurement of Waste," by A. C. Harberger, *Amer. Econ. Rev. Proc.,* May 1964, *54,* 93–96.

R. W. **Shephard,** *Theory of Cost and Production Function,* Princeton 1970.

H. **Theil,** *Economics and Information Theory,* Amsterdam 1967.

———, "On the Geometry and the Numerical Approximation of Cost of Living and Real Income Indices," *De Economist,* 1968, *11,* 677–89.

A. **Wald,** "A New Formula for the Index of the Cost of Living," *Econometrica,* Oct. 1939, *7,* 319–31.

H. **Wold,** "A Synthesis of Pure Demand Analysis, Parts I and II," *Scandinavisk Aktuarietidskrift,* 1943, *26,* 85–144 and 220–272.

———, "A Synthesis of Pure Demand Analysis, Part III," *Skandinavisk Aktuarietidskrift,* 1944, *27,* 69–120.

# The Determinants of Investment in Petroleum Reserves and Their Implications for Public Policy

*By* JAMES C. COX AND ARTHUR W. WRIGHT*

Interest in the determinants of investment in crude oil and natural gas reserves derives from three sources. First, it is always interesting to find a satisfactory explanation of investment behavior in any industry. Second, an aspect of the current concern with the "energy crisis" is the domestic crude petroleum industry's productive capacity, which is an increasing function of the stock of proved oil and gas reserves. Third, there is a decades-old controversy over the special provisions of the federal corporation income tax law which apply to petroleum producers; those special provisions have traditionally been justified by an asserted need to increase investment in petroleum reserves in order to protect "national security."[1]

In three previous papers,[2] we explored the running controversy over the special petroleum tax provisions. In our 1973a paper, we outlined a framework for determining whether the special provisions were cost effective compared to alternative policies for increasing investment in petroleum reserves.[3] We concluded that it was impossible to evaluate the policy because there were as yet no reliable estimates of the determinants of investment in petroleum reserves.

In this paper we present a model of investment in proved reserves in the *U.S.* crude petroleum producing industry, and empirical results for a subsector of that industry for 1959–71, using estimating equations derived from the model. The subsector consists of the five major petroleum producing states which practice "market-demand prorationing."[4] The empirical results indicate that investment in petroleum reserves depended on three public policies: the special federal tax provisions, state market-demand prorationing, and the federal oil import quota. It is possible to draw some tentative policy conclusions from our empirical results, al-

* Associate professors of economics, University of Massachusetts. An earlier draft of this paper was presented at the 1973 Winter Meetings of the Econometric Society. We want to thank Ronald Ehrenberg, Robert Hall, and James Kindahl for helpful advice and criticism. Valuable programming assistance was provided by Donald Barnett, Jr. Partial financial support was provided by the Ford Foundation and by the University of Massachusetts. Neither institution is responsible for the contents of this paper.

[1] The special federal tax provisions for crude petroleum producers are the option to claim "percentage" rather than "cost" depletion on producing wells, and the option to expense rather than depreciate so-called "intangible" drilling costs. Compared to uniform tax treatment of corporate income in all industries, the special provisions for petroleum are a subsidy administered through the revenue side of the budget. Other mineral industries may also claim percentage depletion. See Susan Agria for a detailed discussion of the special tax provisions for mineral industries.

[2] The authors (1973a, b), and Wright.

[3] Even a finding that the tax-subsidy policy was cost-effective in increasing petroleum reserve investment would not, of course, necessarily imply that the policy is "in the public interest," since tradeoffs with other programs would not be considered. See the authors (1975a) for a preliminary application of the cost-effectiveness approach to evaluating the special tax provisions.

[4] The five states are Kansas, Louisiana, New Mexico, Oklahoma, and Texas. According to American Petroleum Institute (*API*) et al. (1972), they contain about 75 percent of *U.S.* proved reserves outside Alaska. The operation of market-demand prorationing is discussed in detail in Sections I and IV below.

though a complete analysis of the three public policies will require additional empirical estimates.

The model is developed in Section I. The data sources and methodology used to outfit the model for empirical testing are discussed in Section II. The empirical results are reported in Section III. Finally, Section IV contains the conclusions, including the policy implications of our results.

## I. A Model of the Crude Petroleum Producer[5]

We assume that crude petroleum producers maximize the present value of after-tax cash flow, subject to the constraints of a production function and an accounting identity. The production function is assumed to be *CES*. The accounting identity relates changes in petroleum reserve stocks to flows of gross investment in reserves and current output ("depreciation").

The after-tax cash flow of a crude petroleum producer is conveniently represented as the difference between a revenue term and two cost terms, one for investment (reserve acquisition) costs and one for other input costs. Thus after-tax cash flow at time t can be written as

$$N(\mathrm{t}) = N_1(\mathrm{t}) - N_2(\mathrm{t}) - N_3(\mathrm{t}) \tag{1}$$

where $N_1$ is the revenue term and $N_2$ and $N_3$ are the investment and noninvestment cost terms, respectively. We discuss each of these cash-flow components in turn.

A crude petroleum producer must make royalty payments to the landowners from whom the drilling and production rights have been leased. These payments are customarily calculated as a percentage of gross revenue.[6] Let $(1-\pi)$ be the proportion of gross revenue which must be paid in royalty; then the proportion $\pi$ of gross revenue accrues to the producer. If $p$ is the price and $Q$ the quantity of marketed output, the revenue of a crude petroleum producer before taxes is $\pi p Q$.

The after-tax revenue of a crude petroleum producer depends on several tax provisions. State and local governments assess production and severance taxes on both quantity of production and revenue; these are represented in the after-tax cash flow equation by the average production and severance tax rate $y$. These taxes are deductible from net income subject to the federal corporation income tax, which is assessed at rate $u$. Further deductions from federal tax are allowed for a proportion $z$ of gross revenue through the percentage depletion allowance. Together, the royalty share and the tax provisions determine the after-tax revenue component of the cash flow equation at time t:

$$N_1(\mathrm{t}) = \{1 - y(\mathrm{t}) - u(\mathrm{t}) \cdot [1 - y(\mathrm{t}) - z(\mathrm{t})]\}\pi(\mathrm{t})p(\mathrm{t})Q(\mathrm{t}) \tag{2}$$

Various categories of investment cost in the crude petroleum industry are treated differently under the federal corporation income tax. Define the following terms: $I(\mathrm{t})$ is total reserve acquisition cost at time t; $q_1(I(\mathrm{t}), \mathrm{t})$ is the proportion of $I(\mathrm{t})$ spent on drilling dry holes; $q_2(I(\mathrm{t}), \mathrm{t})$ is the proportion of $I(\mathrm{t})$ spent on "intangible" costs of successful wells; $q_3(I(\mathrm{t}), \mathrm{t})$ is the proportion of $I(\mathrm{t})$ spent on "tangible" costs of successful wells; and $D(\mathrm{t})$ is the discounted value at time t of the time stream of tax deductions from one dollar of depreciable outlays made at time t. The three proportions $q_j$ sum to one at any time t.

Dry-hole costs $q_1 I$ and intangible costs

[5] In constructing the model, we have benefitted from the earlier work on investment in manufacturing by Dale Jorgenson (1965, 1967) and by Robert Hall and Jorgenson. Throughout the present paper, "petroleum" refers to crude oil, natural gas, and natural gas liquids.

[6] Stephen McDonald (1963, p. 18), explains how the royalty payment comes "off the top" before tax liability is calculated.

$q_2I$ are fully deductible from gross income in the year in which they are incurred. Tangible costs $q_3I$ must be capitalized and depreciated over a number of years; therefore the time t value of the tax deductions they provide is $Dq_3I$. Thus we have the after-tax investment cost component of the cash-flow equation at time t:

$$(3)\quad N_2(t) = \{[1-u(t)][q_1(I(t),t)+q_2(I(t),t)] + [1-u(t)D(t)]q_3(I(t),t)\}I(t)$$

Other input cost categories are also treated differently under the federal corporation income tax. Define the following terms: $L$ is an index of the quantities of nonreserve inputs into the production of crude petroleum; $w_1$ is the expensible cost per unit of $L$; $w_2$ is the depreciable cost per unit of $L$; and $w_3$ is the nondeductible cost per unit of $L$ for producers taking percentage depletion rather than cost depletion. Then the noninvestment cost component of the cash-flow equation at time t is

$$(4)\quad N_3(t) = \{[1-u(t)]w_1(t) + [1-u(t)D(t)]w_2(t) + w_3(t)\}L(t)$$

The quantity of output at time t is constrained by the differentiable implicit production function

$$(5)\quad F(Q(t), \Omega(t), L(t), t) = 0$$

The variable $\Omega$ is the full-time equivalent stock of proved reserves; we specify the proved reserves input that way instead of simply as the stock of proved reserves because of market-demand prorationing (*MDP*). *MDP* limits the use of the productive services of proved petroleum reserves. For example, the Texas Railroad Commission (*TRRC*) formerly imposed "shutdown days" on the operation of wells subject to its control.[7] Thus if there had been fifteen shutdown days in a month of thirty days, the market-demand factor $S$ would have been

$$(6)\quad S = 0.5 = \frac{30-15}{30}$$

If *MDP* were actually enforced as the name shutdown days suggests, by shutting wells down completely part of the time, the full-time equivalent of a stock of reserves $R$ would be simply $SR$; the flow of services from the stock of reserves could then be assumed proportional to $SR$. In fact, *MDP* is enforced differently; a controlled well is permitted to operate every day, so long as total output for the month does not exceed the quantity $S$ times the "rated allowable" capacity of the well (which the *TRRC* also determines). Producers may therefore choose to obtain a given flow of productive services from fewer proved reserves than if they were forced to shut down part of the time; if so, they will utilize their reserves more intensively than they would under a literal shutdown days scheme. To include this possibility, we write $\Omega$ as the function

$$(7)\quad \Omega(t) = S(t)^{\theta}R(t), \qquad 0 \leqq S(t) \leqq 1, \quad 0 < \theta$$

where $\theta$ is the elasticity of the full-time equivalent stock of reserves with respect to the market-demand factor $S$. A literal shutdown days scheme would be the special case of (7) where $\theta$ equals 1. If $\theta$ is less than 1, it would mean that prorationed producers hold a smaller stock of reserves for any flow of productive services under actual *MDP* than under literal shutdown days. A value of $\theta$ greater than 1 would have the opposite implication; this case seems unlikely, but the actual value of $\theta$ is, of course, an empirical question.

Gross additions to proved reserves at time t are represented by the differentiable function $\phi(I(t), t)$. The function $\phi$ is as-

[7] In 1963, the *TRRC* ceased using shutdown days and began setting "market demand factors" directly in percentage terms, as in statement (6). As the text indicates, the two regulatory procedures are equivalent.

sumed to be increasing and strictly concave in $I$ (expenditures on acquiring reserves) for all values of t. This representation assumes that at every point in time, as the size of the reserve acquisition program increases, the marginal addition to reserves per dollar spent decreases. This might occur, for example, if the proportion of "dry holes" increased or if the average quantity of proved reserves per successful well diminished.

Let $r(t)$ be the time-rate of change of the stock of proved reserves, $dR(t)/dt$. Then we have the accounting relation (capital stock equation of motion)

$$r(t) = \phi(I(t), t) - Q(t) \tag{8}$$

That is, the rate of change of the stock of proved reserves at time t must equal the rate of gross additions to reserves at time t less the rate of output from reserves at time t.

The assumed objective of a petroleum producer is maximization of the present value of after-tax cash flow,

$$V = \int_0^\infty N(t) e^{-\int_0^t i(s)ds} dt \tag{9}$$

where $i(s)$ is the after-tax rate of interest. The solution to the problem of maximizing (9), subject to (5), (7), and (8), can be found by maximizing the Lagrangian function

$$\int_0^\infty g(t)dt = \tag{10}$$

$$\int_0^\infty \{N(t)+\lambda(t)F(Q(t), S(t)^\theta R(t), L(t), t)$$

$$+\eta(t)[\phi(I(t), t)-Q(t)-r(t)]\} e^{-\int_0^t i(s)ds} dt$$

Substituting (1)–(4) into (10), we find the Euler necessary conditions to be equations (5) and (8), plus the following:

$$0 = \frac{\partial g(t)}{\partial Q(t)} = \lambda(t)\frac{\partial F}{\partial Q(t)} + \{1 - y(t) - u(t) \cdot [1 - y(t) - z(t)]\}\pi(t)p(t) - \eta(t) \tag{11}$$

$$0 = \frac{\partial g(t)}{\partial L(t)} = \lambda(t)\frac{\partial F}{\partial L(t)} - \{[1 - u(t)]w_1(t) + [1 - u(t)D(t)]w_2(t) + w_3(t)\} \tag{12}$$

$$\begin{aligned} 0 = \frac{\partial g(t)}{\partial I(t)} = \eta(t)\frac{\partial \phi}{\partial I(t)} &- \{[1-u(t)][q_1(I(t), t)+q_2(I(t), t)] \\ &+ [1 - u(t)D(t)]q_3(I(t), t)\} \\ &- \left\{[1 - u(t)]\left[\frac{\partial q_1}{\partial I(t)} + \frac{\partial q_2}{\partial I(t)}\right] \right. \\ &\left. + [1 - u(t)D(t)]\frac{\partial q_3}{\partial I(t)}\right\} I(t) \end{aligned} \tag{13}$$

$$\begin{aligned} 0 &= \frac{\partial g(t)}{\partial R(t)} - \frac{d}{dt}\left(\frac{\partial g(t)}{\partial r(t)}\right) \\ &= \lambda(t)S(t)^\theta \frac{\partial F}{\partial \Omega(t)} + \frac{d\eta(t)}{dt} - i(t)\eta(t) \end{aligned} \tag{14}$$

The preceding necessary conditions can be used to derive the investment functions implicit in the model. We first derive and interpret one argument of those functions, namely, the output-input, after-tax relative price variable. Then we assume a specific form of the production function to derive the investment functions; their discrete approximations comprise the estimating equations used in the empirical work.

The first step is to show that the necessary conditions imply that a petroleum producer should set the marginal product of its stock of reserves equal to the ratio of the marginal after-tax cost of holding reserves to the marginal after-tax net return from producing reserves. Using equations (5) and (7) and the implicit function theorem, we find

$$(15)\quad \frac{\partial Q(t)}{\partial R(t)} = -S(t)^{\theta}\frac{\lambda(t)\partial F/\partial\Omega(t)}{\lambda(t)\partial F/\partial Q(t)}$$

Equations (11) and (14) imply

$$(16)\quad -S(t)^{\theta}\frac{\lambda(t)\partial F/\partial\Omega(t)}{\lambda(t)\partial F/\partial Q(t)} = \frac{i(t)\eta(t) - d\eta(t)/dt}{\{1-y(t)-u(t)[1-y(t)-z(t)]\}\pi(t)p(t)-\eta(t)}$$

Equations (15) and (16) imply

$$(17)\quad \frac{\partial Q(t)}{\partial R(t)} = \left[\left\{i(t) - \frac{1}{\eta(t)}\frac{d\eta(t)}{dt}\right\}\eta(t)\right] \div [\{1-y(t)-u(t)[1-y(t)-z(t)]\}\pi(t)p(t)-\eta(t)]$$

The left-hand side of (17) is obviously the marginal product of the stock of reserves. We now show that the right-hand side of (17) is the ratio of the marginal after-tax net cost of holding reserves to the marginal after-tax net return from producing reserves. Consider first the numerator. From (13), we see that $\eta(t)$ is the marginal after-tax cost of a unit of proved reserves;[8] therefore $[1/\eta(t)][d\eta(t)/dt]$ is the own-rate of interest on reserves. Since $i(t)$ is the after-tax monetary rate of interest, the numerator of the right-hand side of (17) is the marginal after-tax net cost of holding a unit of reserves. Next consider the denominator. The first term $\{1-y(t)-u(t)\ [1-y(t)-z(t)]\}\pi(t)p(t)$ is the marginal after-tax revenue from selling a unit of output. The second term $\eta(t)$ is the marginal after-tax cost of a unit of reserves to replace that which is produced. The difference between the two terms is the marginal after-tax net return from producing a unit of reserves.

[8] Let $\beta=[(1-u)(q_1+q_2)+(1-uD)q_3]+[(1-u)(\partial q_1/\partial I+\partial q_2/\partial I)+(1-uD)\partial q_3/\partial I]I$ in equation (13). Note that $\beta$ is the marginal change in the time t value of after-tax cash flow with respect to investment expenditure. From (13), $\eta=\beta(\partial\phi/\partial I)^{-1}$. Since $\partial\phi/\partial I$ is the marginal increase in reserves from investment expenditure, $(\partial\phi/\partial I)^{-1}$ is the marginal *before-tax* cost of reserves; hence $\beta(\partial\phi/\partial I)^{-1}$ is the marginal *after-tax* cost of reserves at time t.

The right-hand side of (17) is the inverse of the output-input, after-tax relative price variable $h$, which is shown below to be an argument of the investment functions:

$$(18)\quad h(t) = [\{1-y(t)-u(t)[1-y(t)-z(t)]\}\pi(t)p(t)-\eta(t)] \div \left[\left\{i(t) - \frac{1}{\eta(t)}\frac{d\eta(t)}{dt}\right\}\eta(t)\right]$$

We assume a *CES* production function,

$$(19)\quad Q(t) = Ae^{\gamma t}\{a\Omega(t)^{-v} + (1-a)L(t)^{-v}\}^{-b/v}$$

where $A>0$ is the scale parameter; $\gamma\geqq 0$ is the rate of technological change; $a\in(0,1)$ is the input-intensity parameter; $(1+v)^{-1}\in(0,1]$ is the elasticity of factor substitution, restricted so that both inputs are necessary for positive production; and $b\in(0,1]$ is the degree of homogeneity, restricted so that the production function is concave. Equations (7) and (19) imply that the marginal product of reserves is

$$(20)\quad \frac{\partial Q(t)}{\partial R(t)} = \frac{\partial Q(t)}{\partial\Omega(t)}\frac{\partial\Omega(t)}{\partial R(t)} = abA^{-v/b}[e^{\gamma t}]^{-v/b}[Q(t)]^{(1+v/b)}[R(t)]^{-(1+v)}[S(t)]^{-\theta v}$$

From equations (17), (18), and (20), we obtain the following expression for the optimal stock of proved reserves:

$$(21)\quad R(t) = [abA^{-v/b}]^{1/(1+v)}\cdot[h(t)]^{1/(1+v)}[Q(t)]^{(b+v)/b(1+v)}\cdot[S(t)]^{-\theta v/(1+v)}[e^{\gamma t}]^{-v/b(1+v)}$$

Taking a logarithmic transformation of (21) yields the optimal reserves stock equation

$$(22)\quad ln\,R(t) = \alpha_0 + \alpha_1\,ln\,h(t) + \alpha_2\,ln\,Q(t) + \alpha_3\,ln\,S(t) + \alpha_4 t$$

where:

$$(23) \quad \alpha_0 = \frac{1}{1+v}[ln\ a + ln\ b] - \frac{v}{b(1+v)} ln\ A$$

$$\alpha_1 = \frac{1}{1+v}$$

$$\alpha_2 = \frac{b+v}{b(1+v)}$$

$$\alpha_3 = -\frac{\theta v}{1+v}$$

$$\alpha_4 = -\frac{\gamma v}{b(1+v)}$$

Replacing continuous time by discrete time in (22), and adding the error term $\epsilon_t$, we get the reserves stock estimating equation

$$(24) \quad ln\ R_t = \alpha_0 + \alpha_1\ ln\ h_t + \alpha_2\ ln\ Q_t + \alpha_3\ ln\ S_t + \alpha_4 t + \epsilon_t$$

Differentiation of (22) yields the proportional net investment equation

$$(25) \quad \frac{dR(t)/dt}{R(t)} = \alpha_4 + \alpha_1 \frac{dh(t)/dt}{h(t)} + \alpha_2 \frac{dQ(t)/dt}{Q(t)} + \alpha_3 \frac{dS(t)/dt}{S(t)}$$

Replacing continuous time by discrete time in (25), and adding the error term $\epsilon_t'$, we get the proportional net investment estimating equation

$$(26) \quad \frac{\Delta R_t}{R_t} = \alpha_4 + \alpha_1 \frac{\Delta h_t}{h_t} + \alpha_2 \frac{\Delta Q_t}{Q_t} + \alpha_3 \frac{\Delta S_t}{S_t} + \epsilon_t'$$

The net investment estimating equation with error term $\epsilon_t''$ is

$$(27) \quad \Delta\ ln\ R_t = \alpha_4 + \alpha_1 \Delta ln\ h_t + \alpha_2 \Delta\ ln\ Q_t + \alpha_3 \Delta\ ln\ S_t + \epsilon_t''$$

Equation (27) can be derived either by taking a first-order Taylor series approximation of (22), or by taking first differences in (24).

## II. Data Sources

In this section, we discuss the selection and use of empirical measures of the cost of acquiring proved reserves and of the output and price of crude petroleum. We also briefly describe the other data used in the estimations.

### A. *Cost of Acquiring Petroleum Reserves*

The only suitable data on reserve acquisition costs which are broken down by state are those published by the Joint Association Survey (*JAS*) for "costs of drilling and equipping wells" (hereafter *D&E* costs). These data are available separately for successful oil wells, successful gas wells, and total dry holes; they appear to cover the vast bulk of "proving up" outlays. The *JAS* series on *D&E* costs is only available continuously for the years 1959–71, restricting the empirical estimations to thirteen observations.[9]

### B. *Current Petroleum Output and Price*

For the output variable $Q$, U.S. Bureau

[9] American Petroleum Institute (*API*) et al., Section I. The *D&E* costs include only the "Christmas tree" on wells to be used in production. The procedures underlying the *JAS* data have been criticized by Franklin Fisher; we have not evaluated the data published since Fisher wrote (i.e., those for 1961–71), but the methodology appears to have been substantially improved (see, e.g., Morris Adelman, p. 121). *JAS* published *D&E* costs for 1953, 1955, and 1956, but not for 1954 or 1957 and 1958; the early data are of much lower quality than those for 1959–71. The Chase Manhattan Bank (*CMB*) and the *JAS* (*API* et al., Section II) publish series on exploration and development outlays but only for the entire United States. The *CMB* series would not be suitable for econometric work because it is intended as information for investors, not as a consistently defined time-series. Moreover, it includes production facilities beyond the "Christmas tree."

of Mines (*USBM*) data were used in constructing a Divisia quantity index[10] of current outputs of oil, natural gas (nonassociated and associated-dissolved), and natural gas liquids—i.e., all petroleum production from which revenue was received.[11] The *USBM* natural gas figure used in the output Divisia index was "marketed output," which is equal to "gross" output from oil and gas wells, less "repressuring" and "losses."

For the price variable $p$, a Divisia price index of oil, natural gas, and natural gas liquids was constructed. For oil and natural gas liquids, *USBM* data on values realized "at the well" (oil) and "at plants" (natural gas liquids), divided by the appropriate *USBM* output figure, were used. For natural gas, the relevant price for decisions on new reserves in year t is the price obtained on new contracts made in that year, not average realized prices which include sales under long-term contracts made in years past; accordingly, a series for "new contract" prices prepared by Foster Associates for the Energy Policy Project was used.

### C. *Other Data Sources*

1) The stock of proved reserves $R$. Annual data on end-of-year proved reserves of oil and natural gas, in American Petroleum Institute, American Gas Association, and Canadian Petroleum Association, were used in a Divisia index of reserves.

2) Market-demand factor $S$. The values of this variable were those set by the *TRRC;* the annual market-demand factor was calculated as a percentage equal to the average of the twelve monthly figures.

3) *U.S.* corporate income tax rate $u$. Data for 1959–69 were obtained from Joseph Pechman, p. 118; the rates for 1970 and 1971 were taken from the U.S. Internal Revenue Code.

4) Percentage depletion rate $z$. The statutory percentage depletion rate was used for want of a time-series of the effective rate; the latter rate would be less than the former because of the net income limitation.

5) Average production and severance tax rate $y$. This rate was calculated from data in *API* et al., Section II, on state and local taxes paid on oil and gas production, divided by the total value of petroleum production.

6) Discounted value at time t of $1 of depreciable cost incurred at time t, $D$. Iterations from 0.4 to 0.8 in increments of 0.1 showed very little variation in the results. We used the conservatively high value of $D=0.8$, which is the approximate value of $1 of depreciable cost over five years at 12 percent by the sum-of-years-digits depreciation method (see Hall and Jorgenson).

7) Interest rate $i$. Two alternative time-series for the after-tax rate of interest were used. (a) The quantity $(1-u)$ times Moody's index of all "industrial" bond yields, for thirty-six bonds, referred to as the "debt" interest rate. (b) The quantity $(1-u)$ times the inverse of Standard and Poor's composite "price-earnings ratio," referred to as the "equity" interest rate.

8) Royalty share $(1-\pi)$. Based on information in McDonald (1971, p. 14, and 1963, p. 103, n. 132), we iterated over values from 0.10 to 0.20. Since the empirical results varied little between iterations, we follow *API* et al., Section II, and report the results for a royalty share of 15 percent.

## III. Empirical Estimations

In this section, we first discuss the construction of the relative price variable $h$ from the data discussed in Section II. We

[10] On Divisia quantity and price indexes, see Herman Wold and Lars Juréen, and Jorgenson and Griliches.

[11] The *USBM* data are given in convenient form through 1969 in *API*. For 1970 and 1971, the *USBM*'s *Mineral Industry Surveys* were used.

then report ordinary least squares estimates of the reserves investment equations derived in Section I.

The first step in constructing the time-series of $h$ was to calculate an average-cost proxy for the marginal after-tax cost of reserves $\eta$ defined by (13). To separate $D\&E$ costs on successful wells into intangible (expensible) and tangible (depreciable) components, we regressed intangibles on total $D\&E$ costs for successful oil wells and for successful gas wells, using *JAS* data for 1959–64. The total figure for each type of well was then multiplied by the fitted slope coefficient and by one minus the coefficient to obtain intangible and tangible costs, respectively.[12]

The time-series of average costs of acquiring reserves by type of tax treatment were then obtained by dividing dry-hole, intangible, and tangible $D\&E$ costs by gross additions to petroleum reserves. Let $\Lambda_t$ be a Divisia quantity index of $(R_t - R_{t-1} + Q_t)$ for crude oil and nonassociated natural gas in year t. Also let $X_{dt}$, $X_{ct}$, and $X_{gt}$ be $D\&E$ costs in year t on dry holes, successful crude oil wells, and successful gas wells, respectively. Finally, let $\beta_c$ and $\beta_g$ be the estimated proportions of intangibles in $D\&E$ costs in year t for successful oil wells and successful gas wells, respectively. Then for year t the average dry-hole cost of new reserves was calculated as $X_{dt}/\Lambda_t$; the average intangible cost of new reserves was $(\beta_c X_{ct} + \beta_g X_{gt})/\Lambda_t$; and the average tangible cost of new reserves was $[(1-\beta_c)X_{ct} + (1-\beta_g)X_{gt}]/\Lambda_t$.

The last step in calculating the average-cost proxy for the time-series of $\eta$ was to multiply the above dry-hole and intangible costs by $(1-u)$ and to multiply the above tangible costs by $(1-uD)$. The time-series of $\eta$ in turn was combined with the production-tax rate $y$, the percentage depletion rate $z$, the royalty share $\pi$, the price index of petroleum production $p$, and the rate of interest $i$, to calculate the time-series of the output-input, after-tax relative price variable $h$, given by (18). There is no time trend in the time-series of $\eta$; the ratio of the variance to the mean of the series is less than 0.2 percent. This allowed us to assume that over the time period covered by our data, the own-rate of interest on proved reserves was zero. Consequently, the term $[1/\eta(t)][d\eta(t)/dt]$ was set equal to zero in calculating the time-series of $h$.

Market price $p$ is a component of $h$, and the quantity of output $Q$ enters the estimating equations as another explanatory variable. During the period covered by our data, authorities in the five prorationing states effectively controlled the price of crude oil through the policy of *MDP* (under the protective cover of the import quota). By varying the market-demand factor, the state authorities in effect selected price-quantity pairs from the domestic demand curve. It is therefore appropriate in the following regressions to assume that market price and output were both exogenously determined variables over the period studied.[13]

The results of least squares regressions for the estimating equations derived in Section I, using the data described in Section II and the procedures detailed above, are reported in Table 1 (debt interest rate) and Table 2 (equity interest rate). Given the small number of degrees of freedom, the results of the least-squares regressions are encouraging. All estimated coefficients of explanatory variables are highly

[12] The *JAS* definitions of intangibles and tangibles correspond closely to those in the federal tax law (e.g., *API* et al., 1964, section I, pp. 7–8). Scatter diagrams of intangibles against total $D\&E$ costs for successful wells indicated very tight linear fits for both oil and gas wells. Least squares regressions gave $\bar{R}^2 > 0.999$; the constant terms were not significant.

[13] For further explanation of this point, see fn. 17 and the accompanying discussion in Section IV below.

TABLE 1—REGRESSION RESULTS: DEBT INTEREST RATE

| | Coefficients | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | $h_t$ | $Q_t$ | $S_t$ | $t$ | $\bar{R}^2$ | $DW$ | $F$ |
| | $\alpha_0$ | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\alpha_4$ | | | |
| (24) $\ln R_t$ | 4.3423 | .0335 | .8667 | −.1719 | −.0275 | .9734 | 2.0823 | 110.7915 |
| | (2.99) | (4.18) | (9.40) | (16.92) | (7.17) | | | |
| (26) $\Delta R_t/R_t$ | — | .0443 | .9081 | −.1877 | −.0292 | .8919 | 2.1853 | 31.2575 |
| | | (5.11) | (7.12) | (8.62) | (6.06) | | | |
| (27) $\Delta \ln R_t$ | — | .0379 | .8488 | −.1699 | −.0266 | .8507 | 2.1009 | 21.8947 |
| | | (4.16) | (5.96) | (7.17) | (4.84) | | | |

*Note:* Absolute values of $t$-ratios in parentheses. $\bar{R}^2$ = coefficient of determination adjusted for degrees of freedom. $DW$ = Durbin-Watson statistic; null hypothesis of no serial correlation, $DW=2.0$. $F$ = $F$-statistic; $F(4, 8)$ for equation (24) and $F(3, 8)$ for equations (26) and (27).

TABLE 2—REGRESSION RESULTS: EQUITY INTEREST RATE

| | Coefficients | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | $h_t$ | $Q_t$ | $S_t$ | $t$ | $\bar{R}^2$ | $DW$ | $F$ |
| | $\alpha_0$ | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\alpha_4$ | | | |
| (24) $\ln R_t$ | 3.6807 | .0332 | .9078 | −.1893 | −.0295 | .9723 | 2.4401 | 106.1457 |
| | (2.55) | (4.06) | (9.93) | (18.42) | (7.89) | | | |
| (26) $\Delta R_t/R_t$ | — | .0299 | .9644 | −.1898 | −.0338 | .8417 | 2.3299 | 20.5016 |
| | | (3.91) | (6.30) | (7.10) | (6.04) | | | |
| (27) $\Delta \ln R_t$ | — | .0297 | .9176 | −.1789 | −.0308 | .8185 | 2.5032 | 17.5367 |
| | | (3.58) | (5.91) | (6.60) | (5.32) | | | |

*Note:* See Table 1 for explanation of terms.

significant; moreover, they are quite stable between the stock and the flow estimating equations. The $\bar{R}^2$ are relatively high; indeed, we expected worse fits for the first-difference equations, (26) and (27). The Durbin-Watson statistics evidence virtually no serial correlation of the residuals for the debt interest rate and only a weak tendency towards negative serial correlation with the equity interest rate.[14] The $F$-statistics for all six equations estimated exceed the confidence values at the 1 percent level of significance.

The signs of the estimated coefficients reported in Tables 1 and 2 can be interpreted as follows. For the stock equation (24), producers held a larger stock of proved reserves during the period studied, *ceteris paribus*, (a) the higher was the output-input relative price variable ($\alpha_1>0$); (b) the higher was the rate of current petroleum output ($\alpha_2>0$); (c) the smaller was the market-demand factor ($\alpha_3<0$);[15]

[14] Our colleague, Ronald Ehrenberg, has pointed out that regardless of the calculated values of the Durbin-Watson statistics for the equations estimated, the apparent absence of serially correlated residuals can be true for only the stock equation or the flow equations but not for both simultaneously.

[15] One might be tempted to think that a more stringent (smaller) value of the market-demand factor would have *reduced* the stock of reserves by raising the price of the effective flow of reserve services into the production of crude petroleum (e.g., Adelman, p. 106). In fact, producers were forced by the policy of *MDP* to increase the reserve/output ratio for any given level of output. Hence the negative relationship which we find is the one to be expected.

TABLE 3—IMPLIED PARAMETER VALUES: DEBT INTEREST RATE

| Equations | Parameters $1/(1+v)$ | $b$ | $\theta$ | $\gamma$ |
|---|---|---|---|---|
| (24) $\ln R_t$ | .0335 | 1.1599 | .1779 | .0330 |
| (26) $\Delta R_t/R_t$ | .0443 | 1.1064 | .1964 | .0338 |
| (27) $\Delta \ln R_t$ | .0379 | 1.1865 | .1766 | .0327 |

*Note:* We lack information to solve explicitly for $A$ (scale parameter) and $a$ (input-intensity parameter); see (23).

TABLE 4—IMPLIED PARAMETER VALUES: EQUITY INTEREST RATE

| Equations | Parameters $1/(1+v)$ | $b$ | $\theta$ | $\gamma$ |
|---|---|---|---|---|
| (24) $\ln R_t$ | .0332 | 1.1054 | .1958 | .0337 |
| (26) $\Delta R_t/R_t$ | .0299 | 1.0381 | .1957 | .0362 |
| (27) $\Delta \ln R_t$ | .0297 | 1.0928 | .1844 | .0347 |

*Note:* See Table 3 for explanation.

and (d) the lower was the rate of technological change ($\alpha_4<0$). For analogous interpretations of the two flow equations (26) and (27), add a modifying phrase, "proportional change in" or "change in," where appropriate in the above interpretation of the stock equation.

We report in Tables 3 and 4 the values of the production function parameters $1/(1+v)$, $b$, $\theta$, and $\gamma$ which from (23) are implied by the estimated coefficients reported in Tables 1 and 2. The estimates of $1/(1+v)$, which is the elasticity of substitution between reserves and nonreserve inputs and is equal to the coefficient $\alpha_1$, are contained in $(0, 1]$ as required by our specification of the production function. In addition, the values of the elasticity of substitution are quite small, plausibly suggesting that it is difficult to substitute nonreserve inputs for proved reserves in producing crude petroleum. The estimates of $b$, $\theta$, and $\gamma$ are biased because they are non-linear functions of the $\alpha_j$. The estimates of $b$, which is the homogeneity parameter and is equal to $(\alpha_1-1)/(\alpha_1-\alpha_2)$, are not contained in $(0, 1]$ as required by our specification of the production function. The estimates of $\theta$, which is the elasticity of the full-time equivalent stock of reserves with respect to the market-demand factor and is equal to $\alpha_3/(\alpha_1-1)$, are positive as required by (7); they are also much smaller than unity, suggesting that crude petroleum producers in the prorationing sector did indeed respond to *MDP* by utilizing their proved reserves more intensively than they would have under literal "shutdown days" prorationing. Finally, the estimates of $\gamma$, which can be interpreted as the rate of technological change and is equal to $\alpha_4/(\alpha_1-\alpha_2)$, are nonnegative as required by our specification of the production function; furthermore, the implied values of 3.3 to 3.6 percent per annum are reasonable.

## IV. Conclusions and Policy Implications

Our empirical estimates of the determinants of investment in crude petroleum reserves in the prorationing sector of the *U.S.* petroleum industry for 1959–71 are consistent with the model of investment in crude petroleum reserves presented in Section I. The results indicate that several public policies have significantly affected investment in reserves in the prorationing sector. Three of the four explanatory variables depend on instruments of government control. The relative price variable is a function of the special federal corporation income tax provisions for petroleum. The market-demand factor is under the direct control of state prorationing authorities. In addition, both those authorities and the federal government exercised control over crude oil price and quantity during the period studied, through the policies of *MDP* and the oil import quota, respectively.

Two of the three policies had direct par-

tial effects on investment in petroleum reserves which can be inferred from our estimates. In addition, all three policies had indirect partial effects through induced changes in equilibrium price and quantity, which can also be inferred from our estimates. The total effects can be determined from our results if both the direct and the indirect effects have the same sign; otherwise, further information is required. We consider first the direct and then the indirect policy effects on petroleum reserves.

The special federal tax provisions for petroleum corporation income increase the value of the relative price variable $h$, given the price of petroleum output.[16] Thus the significantly positive estimates of $\alpha_1$ imply that those provisions increased investment in reserves compared to uniform tax treatment. Similarly, the significant positive estimates of $\alpha_3$ imply that setting the market-demand factor $S$ at less than unity increased investment in reserves compared to the absence of effective *MDP* ($S=1$).

The oil import quota indirectly affected investment in prorationing-sector petroleum reserves by restricting the quantity of imports. Given the market-demand factor and the special tax provisions, reduced imports would lead to a higher price in the *U.S.* market; the higher price in turn would induce an increase in quantity supplied in the prorationing sector. Therefore, the significantly positive estimates of $\alpha_1$ and $\alpha_2$, the coefficients on relative price and output, imply that the oil import quota increased investment in petroleum reserves.

The indirect effects on petroleum reserves of the special tax provisions and *MDP* involve shifts in the petroleum supply curve. It can be shown that the model developed in Section I leads to a prorationing-sector supply function relating quantity supplied to the market-demand factor, time, and two output-input relative price variables, our $h$ and another relative price, which we have not explicitly defined, for the nonreserve input. In order to represent the supply function as a family of supply curves in price-quantity space, we define the following variable:

$$(28) \quad c(t)=\frac{i(t)\eta(t)}{\pi(t)\{1-y(t)-u(t)[1-y(t)-z(t)]\}}$$

Given the assumption of a zero own-rate of interest on reserves (see Section III), statements (18) and (28) imply

$$(29) \qquad h(t)=\frac{p(t)}{c(t)}-\frac{1}{i(t)}$$

Then holding constant the price of the nonreserve input and the rate of interest, quantity supplied in the prorationing sector at time t is an increasing function of $p$ and $S$ and a decreasing function of $c$. A similar argument leads to a supply function for the nonprorationing sector which is increasing in $p$ and decreasing in $c$; the market-demand factor $S$ does not, of course, appear in the nonprorationing-sector supply function. Since total domestic supply is the sum of the quantities supplied by the two sectors, the total supply function can be represented by the family of supply curves $Q(\cdot)$ in the price-quantity space of Figure 1.

Let $D(p)$ in Figure 1 be the demand curve for petroleum in the domestic market. Then suppose that the special tax provisions were made more generous, so that $c$ decreased from (say) $c''$ to $c'$; as a result, the supply curve would shift from $Q(p, c'', S')$ to $Q(p, c', S')$. Given $D(p)$, the shift in supply would in turn reduce equilibrium price from $p^b$ to $p^a$ and increase the

[16] From (18), $h$ is seen to be an increasing function of the rate of percentage depletion $z$. It can be seen from (13) that the expensing of intangible drilling costs transfers marginal acquisition costs in the amount $[q_2+I(\partial q_2/\partial I)]$ from a term with weight $[1-uD]$ to a term with the smaller weight $[1-u]$.

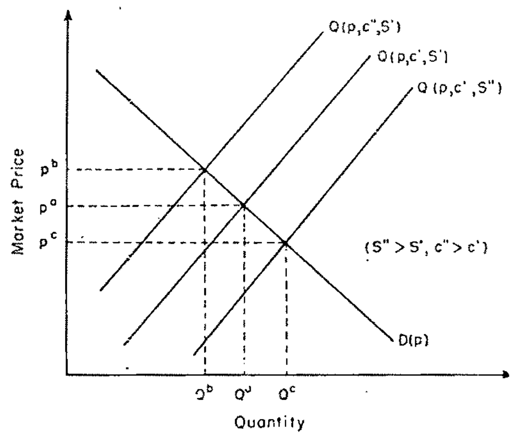



FIGURE 1

quantity supplied from $Q^b$ to $Q^a$. Part of the increase in quantity supplied would come from the prorationing sector. Thus making the special tax provisions more generous would (*ceteris paribus*) reduce market price and increase the quantity supplied (and vice versa for making the provisions less generous).

To isolate the indirect effects of varying the special tax provisions, we take the differential of (22), set $dS_t$ and $dt$ equal to zero, and use (29) to obtain

$$(30) \quad dR_t = -\frac{\alpha_1 p_t R_t}{(c_t)^2 h_t} dc_t + \frac{\alpha_2 R_t}{p_t}\left[\frac{\alpha_1 p_t}{\alpha_2 c_t h_t} - E_t\right] dp_t$$

where $E_t$ is the price elasticity of demand for petroleum at time t. The first term on the right-hand side of (30) is the direct effect on reserves of changes in the special tax provisions; since $dc_t$ is negative, the sign of this term is positive, which is consistent with our earlier finding that the direct effect of the special tax provisions was to increase reserves.

The second expression on the right-hand side of (30) is the indirect effect on reserves of changing the special tax provisions. As noted above in discussing Figure 1, making those provisions more generous reduces the market price. Therefore $dp_t$ is negative and, since $\alpha_2$ is positive, the sign of the indirect effect is the opposite of the sign of the term in brackets. The latter sign is analytically indeterminate, depending on whether $\alpha_1 p_t/\alpha_2 c_t h_t$ is greater or less than $E_t$. To place limits on the sign of the bracketed term, we calculated the values of $\alpha_1 p_t/\alpha_2 c_t h_t$, using the estimates of $\alpha_1$ and $\alpha_2$ (see Tables 1 and 2) plus the observed figures for $p_t$, $h_t$, and $c_t$ for 1959–71; these values vary from 0.0366 to 0.0659. Therefore the bracketed expression in (30) was negative in each year between 1959 and 1971 in which the price elasticity of demand for crude petroleum exceeded 0.0659. If the bracketed expression was negative, the indirect effect as a whole would have been positive, thereby reinforcing the direct effect. On the assumption that $E_t$ was larger throughout the period than the very low value of 0.0659, we tentatively conclude that both the direct and the indirect effects of the special tax provisions increased the stock of proved reserves of petroleum.

To examine the indirect effect of market-demand prorationing on reserves, let us begin with the supply curve $Q(p, c', S'')$ in Figure 1. Now suppose *MDP* is made more stringent by a reduction in the market-demand factor from $S''$ to $S'$. We showed above that this reduction would directly increase the stock of petroleum reserves. There would also be an indirect effect, however, since moving from $S''$ to $S'$ would shift the supply curve in Figure 1 from $Q(p, c', S'')$ to $Q(p, c', S')$.[17] As a result, equilibrium price would increase from $p^c$ to $p^a$ and total quantity supplied would decrease from $Q^c$ to $Q^a$. By reasoning

[17] The ability to shift the supply curve in market price-output space is the reason why effective *MDP* enables the prorationing boards to pick points on the demand curve. Thus, so long as prorationing is effective, market price and output are exogenous to the prorationing sector producer.

analogous to that above for the special tax provisions, the indirect effect of *MDP* would be a decrease in reserves, provided that the price elasticity of demand for crude petroleum was greater than 0.0659. Because the direct and indirect effects have opposite signs in this case, the direction of the total effect of *MDP* on petroleum reserves cannot be determined from our estimates alone. Additional information on the supply and demand functions is required to determine the net total effect.

The Arab embargo of October 1973 has prompted interest in the question of national "independence" in oil. Independence can be defined as having the capacity for self-sufficiency—that is, the capability of being independent of foreign suppliers if the need arises. It can be promoted through actual self-sufficiency in production or through holding excess domestic capacity which can be used in the event that foreign supplies are disrupted. For a depletable resource such as petroleum, increasing domestic output in order to pursue self-sufficiency in production in one period will make it more expensive to be self-sufficient or to hold excess capacity in later periods. In contrast, a policy which promotes independence by inducing domestic producers to hold excess capacity need not mortgage future independence by increasing the present rate of depletion of the resource.[18]

An interesting question is whether past public policies have contributed to or detracted from national independence in oil. Proponents of those policies have claimed that they promoted a "strong petroleum industry" and thus increased "national security." Our empirical estimates permit us to shed some light on this question. In what follows, we analyze the effects of past policies on self-sufficiency as measured by the quantity of imports; discussion of policy effects on alternative measures of self-sufficiency is relegated to footnotes. We also analyze the effects of past policies on independence as measured by the ratio of imports to domestic proved reserves; because reserves are a measure of productive capacity, this ratio is one index of the capability of the domestic petroleum industry to replace imports.

We saw above that, *ceteris paribus*,[19] the indirect effects of the special petroleum tax provisions were transmitted in part through an increase in domestic quantity supplied. Given that the oil import quota was administered by limiting imports to a fixed percentage of domestic production,[20] the increase in domestic production led to an increase in the quantity of imports. Thus the special tax provisions tended to reduce self-sufficiency in oil during the period 1959–71.[21] In addition, the increased domestic production caused the faster depletion of domestic petroleum resources; hence the existence of the special tax provisions in the past has made the present pursuit of independence more costly. Finally, we saw above that the total effect of the special tax provisions was probably to increase investment in proved reserves. Since those provisions also increased imports, we cannot determine their net effect on past independence in oil as

[18] These points are fully explained in the authors' paper, 1975b.

[19] All changes and effects discussed below are partial ones.

[20] James Burrows and Thomas Domencich, p. 12; Cabinet Task Force, p. 10. There was a gradual accumulation of exceptions to this policy criterion but they are irrelevant to an analysis of partial effects.

[21] Other measures of self-sufficiency are the ratio of imports to domestic production $Q_m/Q_d$, and the ratio of imports to total domestic quantity demanded $Q_m/D$. Given the way the import quota was administered, the first measure $Q_m/Q_d$ could not be changed by any other petroleum policy. Because the special tax provisions increased both $Q_m$ and $D$, the second measure $Q_m/D$ would have varied with tax policy; unfortunately, our estimates do not enable us to calculate the relative magnitudes of the two increases. Note that a different administration of the oil import quota would have led to other conclusions about tax policy effects on the several measures of self-sufficiency and independence.

measured by the ratio of imports to proved reserves.

As we showed above, the policy of market-demand prorationing (*MDP*) reduced domestic petroleum production. Given the way the import quota was enforced, *MDP* therefore reduced oil imports and thereby increased past energy self-sufficiency.[22] Because the direct effect of *MDP* on reserves was positive but the indirect effect was negative between 1959 and 1971, we cannot determine the net effect of *MDP* on past independence in oil as measured by the ratio of imports to reserves. We can say, however, that past *MDP* made the present pursuit of independence less costly because the lower rate of petroleum production reduced the rate of depletion of domestic petroleum resources.[23]

The oil import quota, of course, reduced the quantity of imports compared to a policy of free trade in oil; thus it increased past self-sufficiency in oil.[24] In addition, we found that the quota indirectly increased investment in proved reserves by raising the market price of petroleum. The combination of lower imports and greater reserves means that the quota increased past independence in oil as measured by the ratio of imports to reserves. The increased market price, however, led to larger domestic production and hence faster depletion of domestic petroleum resources. The oil import quota of the past therefore made the pursuit of present independence more expensive.

A lesson in the importance of evaluating related public policies simultaneously rather than in isolation from one another is provided by the interaction between the import quota and the special petroleum tax provisions. The quota was ostensibly intended to promote self-sufficiency in oil. We saw above, however, that the special tax provisions in the presence of the quota tended to reduce self-sufficiency in oil during the period 1959–71 by increasing the quantity of imports. Ironically, had there been no import quota, the special tax provisions would have *reduced* oil imports,[25] thereby increasing self-sufficiency. Furthermore, without the quota the special tax provisions would have increased past independence in oil, in that the increase in reserves coupled with the reduction in imports would have decreased the import-reserve ratio. With the quota, in contrast, both the numerator and the denominator of the ratio were increased, leaving the effect on independence indeterminate.

In conclusion, we wish to stress two implications of the preceding discussion of independence in oil. First, in evaluating a particular petroleum policy, one must take into account possible interactions with other policies; we found, for example, that the effects of the special tax provisions on self-sufficiency would have been reversed, had the oil import quota not existed. Second, our analysis reveals that past policies, contrary to assertions by their proponents, did not unambiguously promote national

[22] Using the alternative measure $Q_m/D$, the effect of prorationing on self-sufficiency is qualitatively indeterminate, since both components of the ratio would be reduced.

[23] We abstract here from the considerable inefficiencies, due largely to overdrilling and favoritism towards "stripper" wells, introduced by *MDP* when it was an effective policy; see Adelman.

[24] The quota unambiguously increased self-sufficiency according to the alternative measure $Q_m/Q_d$, since the numerator was reduced and the denominator was increased. According to the measure $Q_m/D$, however, the effect is ambiguous, since both components were reduced (domestic quantity demanded was reduced by the higher market price caused by the quota).

[25] If the supply curve of imports was horizontal and (in the absence of an import quota) constituted the controlling marginal supply price in the domestic market, the increase in domestic supply resulting from more generous federal tax treatment would mean larger domestic production and hence fewer imports, since the total quantity demanded would not change. For an upward-sloping import supply curve, the increase in domestic supply would reduce the market price and hence also reduce the quantity of imports supplied in the market.

independence in oil. In designing future petroleum policies, then, attention should be paid to achieving consistency among different policies. Moreover, public officials would do well to explore alternative sets of policies which may dominate the past set in the sense of offering (say) a lower-cost time path for national independence in oil (see, for example, the authors, 1975b).

## REFERENCES


M. Adelman, "Efficiency of Resource Use in Crude Petroleum," *Southern Econ. J.*, Oct. 1964, *31*, 101–22.

S. Agria, "Special Tax Treatment of Mineral Industries," in A. Harberger and F. Bailey, eds., *The Taxation of Income from Capital*, Washington 1969.

J. Burrows and T. Domencich, *An Analysis of the United States Oil Import Quota*, Lexington 1970.

J. Cox and A. Wright, (1973a) "The Economics of the Oil Industry's Tax Burden," in *The Petroleum Industry's Tax Burden*, Arlington 1973.

——— and ———, (1973b) "Federal Tax Policy and Energy Problems," in U.S. Cong., House, Committee on Ways and Means, *General Tax Reform*, Part 9 of 11, Washington 1973.

——— and ———, (1975a) "The Cost-Effectiveness of Federal Tax Subsidies for Petroleum Reserves: Some Empirical Results and Their Implications," in G. Brannon, ed., *Studies in Energy Tax Policy*, Boston 1975.

——— and ———, (1975b) "A Tariff Policy for Independence from Oil Embargoes," *Nat. Tax. J.*, Mar. 1975, *28*, 29–42.

F. Fisher, *Supply and Costs in the U.S. Petroleum Industry: Two Econometric Studies*, Baltimore 1964.

R. Hall and D. Jorgenson, "Tax Policy and Investment Behavior," *Amer. Econ. Rev.*, June 1967, *57*, 391–414.

D. Jorgenson, "Anticipations and Investment Behavior," in J. Duesenberry et al., eds., *The Brookings Quarterly Econometric Model of the United States*, Chicago 1965.

———, "The Theory of Investment Behavior," in R. Ferber, ed., *Determinants of Investment Behavior*, New York 1967.

——— and Z. Griliches, "The Explanation of Productivity Change," *Rev. Econ. Stud.*, July 1967, *34*, 249–83.

S. McDonald, *Federal Tax Treatment of Income from Oil and Gas*, Washington 1963.

———, *Petroleum Conservation in the United States: An Economic Analysis*, Baltimore 1971.

J. Pechman, *Federal Tax Policy*, rev. ed., New York 1971.

H. Wold and L. Juréen, *Demand Analysis*, New York 1935.

A. Wright, "Federal Tax Policy and the Extractive Industries: Are We Getting Our Money's Worth?," U.S. Cong., House, Committee on Ways and Means, *Tax Reform 1969*, Part 9 of 15, Washington 1969.

American Petroleum Institute, *Petroleum Facts and Figures, 1971*, Washington 1971.

——— et al., *Joint Association Survey of the U.S. Oil and Gas Producing Industry*, Sections I and II, data years 1953, 1955–56, 1959–71, Washington 1956–72.

———, American Gas Assn., and Canadian Petroleum Assn., *Reserves of Crude Oil, Natural Gas Liquids, and Natural Gas in the United States and Canada*, Vol. 26, Washington, Arlington, Calgary, May 1972.

Cabinet Task Force on Oil Import Control, *The Oil Import Question*, Washington 1970.

Chase Manhattan Bank, *Capital Investments of the World Petroleum Industry*, New York 1972.

Foster Associates, "Initial Prices and Volumes [of Natural Gas] under New Contracts, 1952–1971," unpublished.

U.S. Bureau of Mines, *Mineral Industry Surveys*, Washington 1969–71.

U.S. Department of Commerce, *Census of Minerals Industries*, Vol. II, Washington 1954, 1958, 1963, 1967.

# Do Income Taxes Reduce the Effectiveness of Monetary Policy?

*By* Stanley W. Black and Rudolph C. Blitz*

The contemporary lore of monetary and fiscal policy contains two general propositions asserting that high income taxes reduce the effectiveness of monetary policy. One is a reduction in the interest sensitivity of investment demand, and the second is a fall in the value of the multiplier. For example, Warren Smith's recent textbook *Macroeconomics* describes the first effect:

> . . . there are a priori reasons for doubting that investment decisions would be likely to be strongly affected by changes in borrowing costs. One is that interest is deductible as an expense in calculating income for tax purposes. With a corporate income tax rate of 50 percent, an increase in the interest rate by one percentage point, from 5 to 6 percent, would mean an increase of only half a percentage point, from 2.5 to 3 percent, in the *effective* interest rate after adjustment for tax deductibility. Thus, the deductibility of interest under the income tax probably serves to blunt rather substantially the effects of interest rate changes on investment. [p. 204]

It is also generally known that income taxes reduce the value of the multiplier, since the taxes are a leakage from the income stream. Therefore, a reduction in the tax rates on personal and corporate income would be expected to have a two-fold effect in strengthening the impact of monetary policy. First, the interest sensitivity of investment spending would be increased so that a fall in the interest rate would have a larger impact on investment demand. Second, the multiplier would be larger at a lower tax rate so that the larger increment in investment demand would have a magnified effect on aggregate income. Here we assume that the personal and corporate tax rates are changed together in an across-the-board fashion.

In the course of investigating these two propositions rather carefully, we have discovered that the first one is incorrect for business investment, at least in the context of two well-known theories of investment behavior and the present tax structure of the United States. This was earlier pointed out by Paul Samuelson in a brief remark which did not examine the full implications of tax rates for investment. In fact, no general statement about the effects of tax rates on the interest sensitivity of investment can be made without considering the prevailing mix of debt and equity (in turn, influenced by the tax system) and the relative flows of investment to sectors subject to different tax rules (the business and household sectors). This paper is devoted to the exposition of these findings. To begin with, we discuss the effects of tax rates on the cost of capital as defined in the SSRC-MIT-PENN quarterly econometric model of the *U.S.* economy (the *SMP* model) and the underlying Jorgenson and Modigliani-Miller theories of investment behavior. In this context, we demonstrate how the tax structure combined with the *SMP* specification of cost of capital invalidates the first proposition on the sensitivity of investment to the after-tax rate of interest. Next, we illustrate our findings in a simple *IS-LM* framework, showing how changes in tax rates in turn change the effectiveness of monetary policy, both through the interest sensitivity of investment and through the multiplier.

## I. The Effect of Tax Rates on the Interest Sensitivity of Investment

At the most elementary level, the cost of capital for a firm is the opportunity cost of the debt or equity funds which must be attracted to permit the acquisition of physi-

* Vanderbilt University.

cal assets. The rental price of a capital good can be expressed as the quasi rent $Q$ which must be earned by the marginal capital good to justify its purchase, given the price of the capital good $P$, the marginal cost of capital as a percentage rate $R$, and the tax structure and other costs. Ignoring taxes and depreciation for a moment, the marginal investment must satisfy $Q/P=R$. A general formula for the rental price based on Robert Hall and Dale Jorgenson's theory of investment behavior has been developed by Frank de Leeuw and Edward Gramlich to describe the investment functions of the *SMP* model.

Let $D$ be the physical rate of depreciation, $TP$ be the property tax rate, and $G$ the expected rate of capital gains. Denote the income tax rate by $t$, the rate of investment tax credit by $u$ (treated as a reduction in the cost of purchasing the capital good). Let $v$ and $z$ be the proportions of the rental $Q$ and the capital gain $G$ that are taxable, while $w$, $x$, and $y$ are the proportions of depreciation $D$, cost of capital $R$, and property tax $TP$ that are deductible from taxable income. Then the marginal investment must earn a quasi rent given by the following formula:

$$\text{(1)} \quad \frac{Q(1-vt)}{P(1-u)} = D(1-wt) + R(1-xt) + TP(1-yt) - G(1-zt)$$

Under stationary assumptions and perfect certainty, this formula can be derived very easily.[1]

To draw out the implications of equation (1) for our problem, let us ignore depreciation, property taxes, expected capital gains, and the investment tax credit. Then one has

$$\text{(2)} \quad \frac{Q}{P} = R\frac{(1-xt)}{(1-vt)}$$

$$\text{(3)} \quad \frac{\partial Q/P}{\partial R} = \frac{1-xt}{1-vt}$$

Now for fully mortgaged owner-occupied residential housing or other credit-financed consumer durables, the rental is nontaxable ($v=0$), but the interest is deductible ($x=1$). The result is $Q/P=R(1-t)$, and a reduction in the tax rate raises the derivative of the rental rate with respect to the cost of capital (since $\delta(Q/P)/\delta R=1-t$ in this case). This is entirely consistent with the first proposition discussed above.

In the case of business investment, however, all income is taxable ($v=1$) while only a part of the cost of capital, that relating to debt-financed investment, is deductible, so that $x$ is substantially less than one. In the case of all-equity finance, $x=0$. For the *SMP* model, Charles Bischoff took $x=.2$ based on the ratio of debt to equity in the corporate sector. (This assumes that the debt-equity ratio is a constant which may not be true when there is a change in the tax rate. We return to this point in fn. 3.) Therefore we obtain

$$\text{(4)} \quad Q/P = \frac{R(1-.2t)}{1-t}$$

and a cut in the tax rate *lowers* the derivative of the rental rate with respect to the cost of capital! Allowing for depreciation and the investment tax credit makes no difference to this result, which is based directly on the application of Jorgenson's theory of investment behavior. As a matter of fact, exactly the same result as given in (4) can be obtained from Modigliani and Miller's theory of investment behavior under uncertainty.[2]

As should be apparent by now, the actual effect of a change in tax rates on the interest

[1] Ignoring income taxes to begin with, and letting $D'=D\cdot P$, $TP'=TP\cdot P$, $G'=G\cdot P$ denote the *amounts* of depreciation, property taxes, and expected capital gains, one has the condition for profitability of the investment as:

$$P = \sum_{i=1}^{\infty} \frac{Q-D'-TP'+G'}{(1+R)^i} = \frac{Q-D'-TP'+G'}{R}$$

Then

$$R = \frac{Q}{P} - \frac{D'}{P} - \frac{TP'}{P} + \frac{G'}{P} = \frac{Q}{P} - D - TP + G$$

or

$$\frac{Q}{P} = R + D + TP - G$$

Allowing for taxes, one must define each of these factors *after tax*, leading to equation (1).

[2] According to Modigliani and Miller, the cost of capital for a firm is "the minimum prospective yield an investment project must offer to be just worth undertaking from the standpoint of the present stockholders" (1963, p. 439). In their notation, $\overline{X}$ is the

sensitivity of investment will depend crucially on the taxable proportion of the quasi rent of capital $v$ relative to the deductible proportion of the cost of capital $x$, as defined by the tax laws. In the above cases we had $v<x$ or $v>x$. But if $v=x$, the formula again becomes $Q/P=R$, and the tax rate has no effect. Sample situations with $v=x$ include either business investment completely financed by debt ($v=x=1$); or equity-financed outlays for research and development, mineral exploration costs charged to current expense, and advertising outlays to enhance goodwill ($v=x=0$). Richard Musgrave, p. 343, shows that $v=0$ for investment outlays charged to current expense.

The use of equation (3) to analyze the effectiveness of monetary policy depends on the assumption that the after-tax cost of capital, at which the future earnings of investment projects are discounted, is a variable influenced by current market yields. As Bischoff notes, "... the appropriate empirical measure of the discount rate is subject to a great deal of disagreement, but from the point of view of policy the choice is crucial" (p. 83). Hall and Jorgenson, for example, assume that the discount rate or cost of capital is a constant, while Modigliani and Miller (1958) try to show that it is influenced by market interest rates.

## II. An *IS-LM* Framework

For the purpose of illustrating our results on changes in the interest sensitivity of investment caused by changes in the tax rate and integrating them in turn with the effect of tax changes on the value of the multiplier, we shall use a simple linear *IS-LM* framework.

Defining all variables in real terms, let consumption $C$ be a linear function of disposable income $Yd$:

$$C = aY_d + b \tag{5}$$

Assuming a proportional tax rate $t$ and no business saving, disposable income $Y_d$ is a linear function equal to national product $Y$ minus taxes $T$:

$$Y_d = (1-t)Y = Y - T \tag{6}$$

Since disposable income is equal to consumption plus personal saving $S$, we can use (5) and (6) to solve for saving plus taxes as a function of total income.

$$S + T = Y - C = [1 - a(1-t)]Y - b \tag{7}$$

We assume that investment depends linearly on the before-tax required quasi rent on capital $R(1-xt)/(1-vt)$ from equation (2) above:

$$I = g - h\left[\frac{1-xt}{1-vt}\right]R \tag{8}$$

Equating saving plus taxes from equation (7) to investment from equation (8) plus exogenous government spending $\overline{G}$, we find the equation of the $IS$ curve to be

$$R = \frac{(1-vt)}{(1-xt)}\left[\frac{b+g+\overline{G}}{h} - \frac{1-a(1-t)}{h}Y\right] \tag{9}$$

expected earnings before taxes and interest of a firm in some risk class $k$, $\tau$ is the tax rate, $R$ the interest payments on debt $D$ at rate $r$. Then the expected after-tax return to capital is

$$\begin{aligned}\overline{X}^{\tau} &= (1-\tau)(\overline{X}-R) + R \\ &= (1-\tau)\overline{X} + \tau R = (1-\tau)\overline{X} + \tau r D\end{aligned}$$

The value of the firm $V$ is then determined by capitalizing the uncertain stream $(1-\tau)\overline{X}$ at rate $\rho$, "the rate at which the market capitalizes the expected returns net of tax of an unlevered company of size $\overline{X}$ in class $k$," and by capitalizing the certain stream of tax savings on interest payments at rate $r<\rho$, or

$$V = \frac{(1-\tau)\overline{X}}{\rho} + \frac{\tau r D}{r} = \frac{(1-\tau)\overline{X}}{\rho} + \tau D$$

This expression can be manipulated to give the before-tax average cost of capital as

$$\frac{\overline{X}}{V} = \frac{1-\tau D/V}{1-\tau}\rho$$

where $D/V$ is the share of debt in the capital structure of the firm. This before-tax average cost of capital is also the marginal cost of capital $\rho^*$ with which the expected returns on new investments must be compared in making investment decisions, or

$$\rho^* = \frac{1-\tau L^*}{1-\tau}\rho$$

where $L^*$ is the firm's long-run target debt ratio. This is just equation (4) in different notation.

The slope of the *IS* curve is

$$(10)\quad \frac{\partial R}{\partial Y} = -\frac{(1-vt)}{(1-xt)}\left[\frac{1-a(1-t)}{h}\right] < 0$$

A change in the tax rate changes this slope for two reasons, both shown in equation (11): 1) the interest sensitivity of investment changes because of the term $(1-vt)/(1-xt)$, and 2) the multiplier changes because of the term $(1-a(1-t))/h$.

$$(11)\quad \frac{\partial^2 R}{\partial Y \partial t} = \left[\frac{1-a(1-t)}{h(1-xt)^2}\right] \cdot [v(1-xt) - x(1-vt)] - \frac{a(1-vt)}{h(1-xt)}$$

The first term in (11) can be either positive or negative, depending on whether the taxable fraction of profit $v$ is greater or less than the deductible fraction of capital costs $x$, while the second term is always negative.[3]

In Figure 1, the initial equilibrium of the system is given by the solid curves $IS_0$ and $LM_0$, with $IS_0$ derived from the solid curves in the other three quadrants. The effects of monetary policy are shown by a shift from $LM_0$ to $LM_1$ in response to an open market operation.

If the personal and corporate tax rates are reduced, leakages will be lower, as shown by the dashed flatter $S+T$ curve in the southeast quadrant. Thus the multiplier will be larger, and the *IS* curve will become flatter because of this effect, raising the effectiveness of monetary policy. The interest sensitivity of investment shown in the northwest

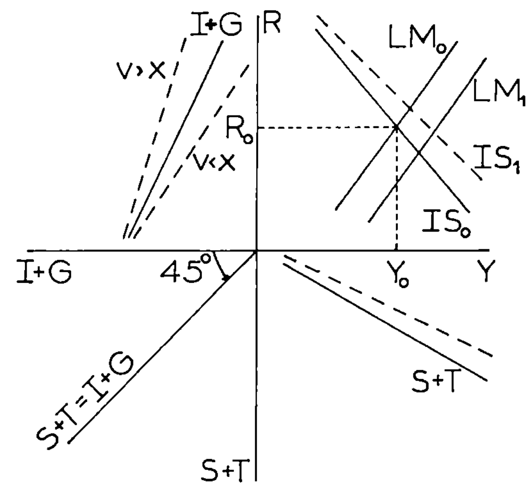



FIGURE 1. EFFECT OF REDUCTION IN RATE OF INCOME TAX

quadrant will either rise or fall, depending on whether $v<x$ (the right-hand dashed curve) or $v>x$ (the left-hand dashed curve). The case $v<x$ makes for a flatter *IS* curve and more effective monetary policy, while the case $v>x$ makes for a steeper *IS* curve and less effective monetary policy. The argument following equation (3) above indicated that $v<x$ for owner-occupied residential housing and other credit-financed consumer durables. On the other hand, $v>x$ for business investment. If $v=x$, the investment function does not shift with the changes in the tax rate.

Presumably, the combined effect of these shifts in the savings and investment functions will be a flatter *IS* curve shifted to the right, as $IS_1$. The presumption is that expression (11) is negative overall. Nevertheless, the effects of higher interest rates on business investment have been shown to be *greater* with high tax rates than with lower tax rates, contradicting Smith's view cited earlier.[4]

### III. Effects of Inflation

The preceding analysis has implicitly assumed a given price level. With an upward sloping Keynesian supply function, the

[3] This assumes that the debt-equity ratio is independent of the tax rate. If we allow the tax rate to affect the debt-equity ratio ($dx/dt>0$), the first term in (11) becomes

$$\frac{1-a(1-t)}{h}\left[\frac{v-x}{(1-xt)^2} - \frac{t(1-vt)}{(1-xt)^2}\cdot\frac{dx}{dt}\right]$$

The term in brackets is negative for $v<x$ and is positive for $v>x$ if $(v-x)/x(1-vt)>(t/x)(dx/dt)$. Thus a variable debt-equity ratio does not reverse our conclusions for the case $v>x$ unless the elasticity of the debt-equity ratio with respect to the tax rate exceeds 8 in the case of the *SMP* model ($v=1$, $x=.2$, $t=.5$).

[4] We have performed simulations with the *SMP* econometric model confirming this finding in the context of the model. A copy of these results is available to interested readers upon request.

analysis would have to be supplemented with discussion of the effects of higher prices on the results of Figure 1. The Keynes and Pigou effects, for example, would shift the curves $IS_1$ and $LM_1$ somewhat to the left as reduced real balances lead to a tighter money market and increased saving. However, the slopes would still change as shown in the figure.

More interesting possibilities arise if we assume that monetary expansion generates a higher rate of anticipated inflation. In this case the nominal rate of interest will rise above the real rate shown in Figure1. Indeed Jorgen Gelting has recently argued that the nominal rate of interest must rise above the real rate by *more* than the anticipated rate of inflation, because the income tax

> ". . . ignores the distinction between that part of interest payments which compensates for the fall in the value of money, and the part which is true income. If in non-inflationary conditions the equilibrium rate of interest gross of tax is 6 percent and thus net of 50 percent income tax is 3 percent then to obtain the same real rate of interest at 10 percent inflation, the market rate should rise to 26.6 percent, equivalent to 13.3 percent net of tax, which will leave a real rate of return of 3 percent. . . . More generally, assuming the rise in the market rate to compensate fully for inflation, the elasticity of the market rate with regard to the tax rate t will be t/(1−t). [p. 65]

Gelting is discussing the return on financial assets held by individuals who are suffering real capital losses ($G<0$ in equation (1)). In this case property taxes, the investment tax credit, and depreciation are irrelevant so the before tax yield is equal to

$$(12) \quad \frac{Q}{P} = \frac{1 - xt}{1 - vt} R + \frac{1 - zt}{1 - vt}(-G)$$

For personal investments, $x=0$, $v=1$, and $z=0$, giving

$$(13) \quad \frac{Q}{P} = \frac{1}{1 - t} R + \frac{1}{1 - t}(-G)$$

Given the real cost of capital $R$, an increase in the anticipated rate of inflation $(-G)$ will raise $Q/P$ by $1/(1-t)$ times the increase in the rate of inflation. It is obvious from equation (12) that this effect is due to the nondeductibility of the real capital losses caused by inflation. If the deductible fraction of losses $z$ were equal to the taxable fraction of yield $v$, a given rate of inflation would raise the nominal yield $Q/P$ by exactly the rate of inflation, unless the real rate of interest falls.

## IV. Concluding Comment

This investigation has shown that contemporary theories of investment behavior in which the cost of capital is taken to be a function of tax rates and the average cost of debt and equity capital yield new conclusions about the effect of taxes on investment behavior. Specifically, when the fraction of earnings that is taxable exceeds the fraction of the cost of capital that is deductible, an increase in the tax rate *raises* the interest sensitivity of investment demand. Conversely, when the taxable fraction of earnings falls short of the deductible fraction of costs, an increase in the tax rate *lowers* the interest sensitivity of investment demand.

There are two major unresolved issues to which our study points. First, the hypothesis that the rate of discount of future earnings prospects is a variable influenced by current market yields is a crucial hypothesis, and one with which we agree.

Secondly, the assumption that the relevant *marginal* cost of capital can be approximated by a weighted average of the *average* cost of debt and equity capital is equally important and much less acceptable. We are not alone in doubting this assumption, which has been questioned by John Lintner, Luigi Tambini, and Joseph Stiglitz, among others. All of these authors in one way or another raise James Duesenberry's argument that the marginal cost is increasing: "As debt rises relative to earnings the risk premium required to cover the leverage of debt service on earnings fluctuation will increase" (p. 94). Empirical evidence on this issue is provided by both Lintner and Tambini.

The effects of deleting the questionable assumption are unclear. A step in this direc-

tion has been taken by Stiglitz, who integrates the corporate and personal income taxes in a general equilibrium analysis of firms and investors.[5]

[5] Our colleague, Fred Westfield, has suggested another direction in which the analysis could be extended. Since the income tax also affects the factor cost of labor, it would be interesting to examine the substitution of capital for labor based on the production function.

## REFERENCES


**C. W. Bischoff**, "The Effect of Alternative Lag Distributions," in G. Fromm, ed., *Tax Incentives and Capital Spending*, Washington 1971.

**G. F. Break**, "The Incidence and Economic Effects of Taxation," in A. S. Blinder et al., eds., *The Economics of Public Finance*, Washington 1974.

**F. de Leeuw and E. M. Gramlich**, "The Channels of Monetary Policy," *Fed. Res. Bull.*, June 1969, *55*, 472–91.

**J. S. Duesenberry**, *Business Cycles and Economic Growth*, New York 1958.

**J. H. Gelting**, "National Autonomy of Stabilization Policy," in A. Ando et al., eds., *International Aspects of Stabilization Policy*, Boston 1975.

**R. E. Hall and D. W. Jorgenson**, "Application of the Theory of Optimum Capital Accumulation," in G. Fromm, ed., *Tax Incentives and Capital Spending*, Washington 1971.

**J. Lintner**, "Corporation Finance: Risk and Investment," in R. Ferber, ed., *Determinants of Investment Behavior*, New York 1967.

**F. Modigliani and M. H. Miller**, "The Cost of Capital, Corporation Finance, and the Theory of Investment," *Amer. Econ. Rev.*, June 1958, *48*, 261–97.

——— **and** ———, "Corporate Income Taxes and the Cost of Capital: A Correction," *Amer. Econ. Rev.*, June 1963, *53*, 433–43.

**R. A. Musgrave**, *The Theory of Public Finance*, New York 1959.

**P. A. Samuelson**, "Tax Deductibility of Economic Depreciation to Insure Invariant Valuations," *J. Polit. Econ.*, Dec. 1964, *72*, 604–06.

**W. L. Smith**, *Macroeconomics*, Homewood 1970.

**J. E. Stiglitz**, "Taxation, Corporate Financial Policy, and the Cost of Capital," *Publ. Econ.*, Feb. 1973, *2*, 1–34.

**L. Tambini**, "Financial Policy and the Corporation Income Tax," in A. C. Harberger and M. J. Bailey, ed., *The Taxation of Income from Capital*, Washington 1969.

# Separability and Vanishing Externalities

*By* W. David Montgomery*

In one of the most influential papers written on the subject of externalities, Otto Davis and Andrew Whinston argue that corrective taxes and private bargaining are likely to achieve an optimum in the presence of mutual externalities between two firms only when externalities are separable, in the sense that marginal cost is independent of the level of externality. Further analysis of the concept of separability reveals that even this conclusion is too optimistic. I shall argue below that the assumptions needed to make taxes and negotiations work properly rule out the possibility of having externalities in any observable situation.

The crucial step in Davis and Whinston's argument is the proposition that separability "... implies the game theoretic concept of dominance" (p. 247). They contend that when the cost function is separable, the profit-maximizing output of a firm affected by an externality is determined independently of the output choice of the firm causing the externality. Thus, when the process of mutual accommodation to externality is formulated as a noncooperative game, choosing that output level is a dominant strategy for the affected firm. Moreover, when the cost function is nonseparable, each firm must base its strategy on expectations about how the other firm will respond to its choices, and the game may fail to have an equilibrium point.

As stated by Davis and Whinston, the argument that separability of the cost function implies the existence of a dominant strategy is incorrect. A strategy should specify both inputs and outputs. In general, the functional relation between inputs and outputs depends on the level of externality. If the input combination which minimizes the cost to one firm of producing a fixed level of output changes as the output of the other firm changes, then it will not be true that the complete optimal strategy for one firm is independent of the actions of the other. The simplest example is the one input, one output case. Suppose an output $Y^*$ is optimal for firm 1 under all actions of firm 2, but that the input required to obtain this output varies as the output of firm 2 varies. Then firm 1 cannot decide on an input-output pair without knowledge of what firm 2 will do.

Separability of the cost function implies only that the output choice is independent of externality. A further condition is needed to guarantee that input choice is also independent. For input choice to be independent of externality, the marginal productivity of each input must be independent of the level of externality. If this is to be the case, the production function must also be separable. Writers following Davis and Whinston, including James Marchand and Keith Russell, have assumed that separability of the cost function is equivalent to separability of the production function. It will be established that this assumption is false, and that it is only possible to have separability of both functions in a special case. I will give two counter examples.

Let $C(Y_1, Y_2)$ be the cost function of a firm which produces $Y_1$ and suffers an external diseconomy which is a function of $Y_2$. The cost function is dual to a production function $F(X_1, \ldots, X_n, Y_2)$.

*Definition* 1: A cost function $C(Y_1, Y_2)$ is separable if and only if it can be written as $C_1(Y_1)+C_2(Y_2)$.

*Definition* 2: A production function $F(X_1, \ldots, X_n, Y_2)$ is separable if and only if it can be written as $g(X_1, \ldots, X_n) + h(Y_2)$.

* California Institute of Technology. I am indebted to Joel Franklin, David Grether, and James Quirk for substantial mathematical assistance and moral support. Research support was provided by the Environmental Quality Laboratory of the California Institute of Technology.

LEMMA 1: *A cost function is separable if and only if* $\partial^2 C/\partial Y_1 \partial Y_2 = 0$ *everywhere. A production function is separable if and only if* $\partial^2 F/\partial X_i \partial Y_2 = 0$, $i = 1 \ldots n$, *everywhere.*[1]

Consider the separable production function

$$F = X_1^{\alpha} X_2^{\beta} - h(Y_2)$$

where $\alpha+\beta<1$. We find the cost function by solving the cost-minimization problem and using the first-order conditions and the production function to eliminate the inputs from the cost equation. From the first-order conditons we have

$$\frac{W_1}{W_2} = \frac{\alpha X_2}{\beta X_1} \tag{1}$$

Solving for $X_1$ and substituting in the production function gives

$$Y_1 = X_2^{\alpha+\beta}\left(\frac{W_2\alpha}{W_1\beta}\right)^{\alpha} - h(Y_2) \tag{2}$$

Solving (2) for $X_2$, and substituting the resulting expression for $X_2$ in (1) enables us to express $X_1$ and $X_2$ in terms of $Y_1$ and $Y_2$ alone. Substitution in $C = W_1X_1 + W_2X_2$ gives the cost function

$$C = W_2\left(1 + \frac{\alpha}{\beta}\right)\left(\frac{W_2\alpha}{W_1\beta}\right)^{-\alpha/(\alpha+\beta)} \cdot (Y_1 + h(Y_2))^{1/(\alpha+\beta)} \tag{3}$$

Clearly (3) is not separable if $\alpha+\beta\neq 1$.

Now consider the nonseparable production function

$$F = (X_1 - h_1(Y_2))^{\alpha}(X_2 - h_2(Y_2))^{\beta}$$

A mathematical development similar to that of the first example yields the separable cost function

$$C = W_2\left(1 + \frac{\alpha}{\beta}\right)\left(\frac{\alpha}{\beta}\frac{W_2}{W_1}\right)^{-\alpha/(\alpha+\beta)} Y_1^{1/(\alpha+\beta)} + W_1h_1(Y_2) + W_2h_2(Y_2)$$

To find conditions under which the production function and its dual cost function are both separable, we express $\partial^2 C/\partial Y_1\partial Y_2$ in terms of the derivatives of the production function. We adopt the following abbreviations

$$\frac{\partial F}{\partial X_i} = F_i \qquad \frac{\partial F}{\partial Y_2} = F_Y$$

$$\frac{\partial^2 F}{\partial X_i \partial X_j} = F_{ij} \qquad \frac{\partial^2 F}{\partial X_i \partial Y_2} = F_{iY}$$

Let

$$\Delta = \begin{vmatrix} F_{11} \ldots F_{1n} & F_1 \\ \vdots \qquad \vdots & \vdots \\ F_{n1} \ldots F_{nn} & F_n \\ F_1 \ \ldots \ F_n & 0 \end{vmatrix}$$

Further let $\Delta_{ij}$ be the $i$, $j$th cofactor of $\Delta$.

THEOREM 1. *A cost function is separable if and only if it is derived from a production function which satisfies*

$$\sum_i F_{iY}\Delta_{n+1,i} + \Delta_{n+1,n+1}F_Y = 0 \tag{4}$$

*at every point which is a proper cost minimum.*[2]

It follows that the proper statement of the relation between separability and dominance is that separability of the cost function *and* of the production function implies the game-theoretic concept of dominance. Moreover, the two separability hypotheses are satisfied simultaneously only when the private part $g$ of the production function has a specific form.

COROLLARY 1: *If the cost function* $C(Y_1, Y_2)$ *and the production function* $F(X_1, \ldots, X_n, Y_2)$ *are both separable and* $F_Y\neq 0$ *everywhere, then* $|g_{ij}| = 0$, *where* $|g_{ij}|$ *is the Hessian determinant of* $g$.

PROOF:

By separability of the cost function (4) holds. Separability of the production function implies $F_{iY} = 0$ for all $i$. Therefore

[1] Lester Ford, p. 251, proves sufficiency. Necessity is trivial.

[2] For proof, see Appendix.

$$\Delta_{n+1,n+1} = \left| g_{ij} \right| = 0$$

COROLLARY 2: *If the cost function and production function are separable and $F_Y \neq 0$, then marginal cost $\partial C/\partial Y_1$ is constant in $Y_1$ and $Y_2$.*

PROOF:

From the Appendix and Paul Samuelson, p. 67,

$$\frac{\partial^2 C}{\partial Y_1^2} = \frac{\partial \lambda}{\partial Y_1} = \lambda \frac{\Delta_{n+1,n+1}}{\Delta} = 0$$

Constancy of marginal cost has some startling consequences. If no externality were present, then the supply function would be such that for any price of output greater than $\partial C/\partial Y_1$ the firm will produce unbounded output. At any price less than $\partial C/\partial Y_1$, it will produce zero output. Only when price of output equals $\partial C/\partial Y_1$ will the firm produce finite output. But when $C_2(Y_2) \neq 0$ there is a fixed cost, imposed on the firm by externality. Therefore even if price is exactly equal to $\partial C/\partial Y_1$ (which is a constant function of $Y_1$ and independent of $Y_2$), the firm will be losing money and will produce zero output But if price is at all higher than $\partial C/\partial Y_1$, the firm can earn unbounded profits by producing infinite output. Thus no equilibrium involving finite, nonzero output by a firm which suffers from an externality and has separable cost and production functions can exist unless the firm which causes the externality is producing zero output.

That is, when the conditions for the existence of dominant strategies in a two-firm externality game are satisfied, the only equilibrium possible is one in which one of the firms is out of business. It follows that it is impossible ever to observe a firm with separable cost and production functions suffering an externality, since either it or the firm causing the externality will always be driven out of business in equilibrium. If each firm creates an externality affecting the other, then only one can survive in equilibrium, and it will not suffer or cause any externality in equilibrium. Moreover, if the two firms produce the same output from the same input, this equilibrium is a Pareto optimum. Since the marginal costs of all firms are constant by hypothesis, one firm can produce any output as efficiently as many. The fixed cost imposed by externality implies that when more than one firm is in operation, more input is needed to produce a given output than is needed when only one firm operates.

The consequence of this analysis is the intensification of the pessimism expressed by Davis and Whinston regarding the possibility of using either corrective taxes or private bargaining to correct externalities. The only circumstances in which they thought such policies workable define a vacuous case, one in which externalities will never be observed or need to be remedied.

## Appendix

### *Proof of Theorem 1*

By Lemma 1 the cost function is separable if and only if $\partial^2 C/\partial Y_1 \partial Y_2 = 0$. We express $\partial^2 C/\partial Y_1 \partial Y_2$ in terms of the production function as follows. Form the Lagrangian expression

$$L = \sum_i W_i X_i + \lambda(Y_1 - F(X_1 \ldots X_n, Y_2))$$

First-order conditions are

$$W_i - \lambda F_i = 0$$

$$Y_1 - F = 0$$

We perturb the solution by varying $Y_1$ and $Y_2$. Totally differentiating the first-order conditions gives the system of equations

$$\text{(A1)} \quad \begin{bmatrix} F_{11} & \ldots & F_{1n} & F_1 \\ \vdots & & \vdots & \vdots \\ F_{n1} & \ldots & F_{nn} & F_n \\ F_1 & \ldots & F_n & 0 \end{bmatrix} \begin{bmatrix} dX_1 \\ \vdots \\ dX_n \\ d\lambda/\lambda \end{bmatrix} = \begin{bmatrix} \dfrac{dW_1}{\lambda} - F_{1Y} dY_2 \\ \vdots \\ \dfrac{dW_n}{\lambda} - F_{nY} dY_2 \\ dY_1 - F_Y dY_2 \end{bmatrix}$$

Solving for $dX_k$ using Cramer's rule gives

$$dX_k \frac{1}{\Delta}\left\{ \sum_{i=1}^{n}\left[\left(\frac{dW_i}{\lambda} - F_{iY}dY_2\right)\Delta_{ik}\right] + (dY_1 - F_Y dY_2)\Delta_{n+1,k}\right\}$$

We assume that $F$ is strictly quasi concave in $X_1 \ldots X_n$, so that $\Delta \neq 0$. Then

$$\frac{\partial X_k}{\partial Y_2} = \frac{-\sum_i F_{iY}\Delta_{ik} - F_Y\Delta_{n+1,k}}{\Delta}$$

Since

$$\frac{\partial C}{\partial Y_2} = \sum_k W_k \frac{\partial X_k}{\partial Y_2} \quad \text{and} \quad W_k = \lambda F_k,$$

$$\frac{\partial C}{\partial Y_2} = \frac{1}{\Delta}\left\{ -\sum_{i=1}^{n}\left[F_{iY}\lambda\left(\sum_k F_k\Delta_{ik}\right)\right] - \lambda F_Y \sum_k F_k \Delta_{n+1,k}\right\}$$

But $\sum_k F_k\Delta_{n+1,k} = \Delta$, and $\sum_k F_k\Delta_{ik} = 0$ since it is an expansion by alien cofactors. Therefore

$$\text{(A2)} \qquad \frac{\partial C}{\partial Y_2} = -\lambda F_Y$$

Differentiating (A2) with respect to $Y_1$ gives

$$\frac{\partial^2 C}{\partial Y_2 \partial Y_1} = -\lambda \sum_i F_{iY}\frac{\partial X_i}{\partial Y_1} - F_Y \frac{\partial \lambda}{\partial Y_1}$$

From (A1),

$$\frac{\partial \lambda}{\partial Y_1} = \lambda\frac{\Delta_{n+1,n+1}}{\Delta}, \qquad \frac{\partial X_i}{\partial Y_1} = \frac{\Delta_{n+1,i}}{\Delta}$$

Therefore

$$\frac{\partial^2 C}{\partial Y_2 \partial Y_1} = -\frac{\lambda}{\Delta}\left(\sum_i F_{iY}\Delta_{n+1,i} + \Delta_{n+1,n+1}F_Y\right)$$

## REFERENCES


O. Davis and A. Whinston, "Externalities, Welfare, and the Theory of Games," *J. Polit. Econ.*, June 1962, *70*, 241–62.

L. Ford, *Differential Equations*, New York 1955.

J. R. Marchand and K. P. Russell, "Externalities, Liability, Separability, and Resource Allocation," *Amer. Econ. Rev.*, Sept. 1973, *63*, 611–20.

P. A. Samuelson, *Foundations of Economic Analysis*, Cambridge, Mass. 1966.

# Estimated Parameters as Dependent Variables

By GARY R. SAXONHOUSE*

Judith Lave and Lester Lave, in their article in this *Review*, have proposed two imaginative, if simple, techniques for estimating cost functions in multiproduct, multiservice industries. Both techniques rest on the assumption that output mix varies among firms but is constant over time within any one firm. The first proposed technique involves two stages of analysis and allows a simple cost function to differ among firms. In the first stage a cost function is estimated for each firm. In the second stage a search is made for the factors causing variation among the estimated individual firm cost function parameters. The second proposed technique is based on the more limited assumption that, apart from the intercept term, all firms have the same cost function. Here the data are pooled and a single cost function estimated.

Lave and Lave, in the course of analysis, conclude that their dichotomous first technique is too inflexible, given their particular problems and opt for the second method. Despite their discarding of the first technique, others have been tempted to try it out and I think some discussion of this proposed estimation procedure is in order.[1] I think this discussion will also have some bearing on the Laves' second method.

The Laves' search for the causes of variation among individual cost functions involves taking firm cost function parameters and regressing them on various sets of independent variables in an effort to find significant relationships. Research workers using such a procedure should be aware that it is unusual in that before these regressions are ever run, an estimate of the variance of each dependent variable is available. Such knowledge may well be unpleasant. If on testing the hypothesis of homogeneity of variance using likelihood ratio methods, the null hypothesis is rejected, then the problem of heteroscedasticity in regressions using estimated parameters as dependent variables must be confronted. The now classical results indicate that applying ordinary least squares to an equation with a heteroscedastic error structure, while giving unbiased estimates of the coefficients of the independent variables, is nonetheless an inefficient method. Also, using the usual formula to calculate sampling variances may well involve an overstatement of the true sampling variances. Given the comparatively low $t$-statistics reported in the Laves' Tables 1, 2, and 3, the possibility exists that many of the variables reported as significant were really not at all significant.

What can be done about this problem of heteroscedasticity? Since estimates of the variance of each observation on the dependent variable are available, a straightforward procedure would imply weighting each observation on all variables used in the second-stage equation by the inverse of the estimated standard error of the dependent variable. This weighting technique, of course, is nothing more than an application of generalized least squares, where an estimate of the variance-covariance matrix is used in place of the true variance-covariance matrix.

One may well wonder whether the two-stage procedure is necessary at all. After all in the Laves' case, the linear second-stage equation (4) can easily be substituted into the linear first-stage equation (3). Unfortunately, to the extent that the variances of the stochastic terms in the first-stage cost equations are hospital specific, the fundamental equation resulting from this substitution will still contain a heteroscedastic error structure. Interestingly enough,

* The University of Michigan. I would like to thank Takeshi Amemiya for good advice and Judith Lave and Lester Lave for providing the data used in the analysis in this paper.

[1] See for example, J. G. Williamson. With increased interest in the last ten years in models where it is assumed that the parameters have a stochastic component, it is not surprising that there also be increased interest in models where the parameters are subject to systematic variation.

efficient estimation of this equation is nothing more than the generalized least squares estimation procedure suggested in the last paragraph![2] Efficient estimation of this equation is the same as applying ordinary least squares to the first-stage equation followed by estimating the second-stage equations using weighted least squares (see the Appendix).

This result is especially interesting since some of the Laves' pooled regressions (especially (4-3) and (4-4)) can be interpreted as resulting from the above suggested substitution. Thus, if it is assumed that the variances of the stochastic terms in these latter regressions are again hospital specific, then efficient estimation will require substantively the same two-stage approach. Efficient estimation in the presence of heteroscedasticity requires modifications in both the Laves' proposed techniques which make them virtually identical.

Table 1 presents the results of the likelihood ratio test on the hypothesis of homogeneity of sampling variance for each of the four estimated first-stage coefficients.[3] In each instance the null hypothesis cannot be accepted. Heteroscedasticity may well be present and it is appropriate to use the weighting procedure outlined above. Thus, each of the Laves' regressions have been run again using this method. The results are presented in Tables 2, 3, and 4. Inspection of the tables indicates the extent to which use of the classical least squares formula to calculate the sampling variance in the presence of the type of heteroscedasticity described above will underestimate the variance. The underestimation varies from none at all to 65 percent. On the average, it is 40 percent. The gains in efficiency which result from the application of the suggested generalized least squares procedure also appear to be substantial. While the actual sampling variances, of course, are not available, the estimated sampling variances under the new procedures are from 40 to 80 percent lower than the correct estimates of the least squares sampling variances. On the average they are 55 percent lower.

TABLE 1—TESTING FOR HOMOGENEITY OF VARIANCE

| Hypothesis[b] | Test Statistic[a] |
|---|---|
| $\sigma^2_{a1T}=\sigma^2_{a2T}=\ldots=\sigma^2_{a74T}$ | $436.352>110.781$ |
| $\sigma^2_{a1U}=\sigma^2_{a2U}=\ldots=\sigma^2_{a74U}$ | $523.187>110.781$ |
| $\sigma^2_{a1F}=\sigma^2_{a74F}=\ldots=\sigma^2_{a74F}$ | $314.586>110.781$ |
| $\sigma^2_{a1S}=\sigma^2_{a2S}=\ldots=\sigma^2_{a74S}$ | [c] $>110.781$ |

[a] $110.781=5$ percent critical value

[b] $\sigma^2_{aij}$=sampling variance on $i$th estimate of $j$th coefficient

[c] Test statistic was too large for the computer to print it out.

While some of the point estimates presented in Tables 2–4 differ substantially from the Laves' estimates, the basic character of the results remains the same. Substantive alteration of their results does occur when regressions having the estimated size coefficient as the dependent variable are run using the new procedure. The new generalized least squares estimates are presented in Table 5. Using ordinary least squares techniques, the Laves are unable to find any variables which significantly contribute to the explanation of the variation of the size coefficient. Similarly, the size coefficient is insignificant in each of the Laves' four pooled regressions. Contrary to both these results,

[2] This result is more than a curiosum. It suggests that attempting to explain the variation among parameter estimates is not the crude *ad hoc* procedure it might appear at first glance. Rather, if appropriate allowance is made for the possible presence of heteroscedasticity (using the weighting procedures outlined above), such regressions are the legitimate second step in an efficient generalized least squares estimation of what could typically be more elaborate models. Given the computational advances of recent years, the literature is full of articles containing many time-series regressions separately estimated under identical specification for firms, industries, states, regions, and nations (or alternatively containing many cross-sections also separately estimated at varying points in time under identical specifications). The above results suggest that efficient systematic explanation of the reported intersample results need not require the underlying raw data (at least this would be true under the specification suggested in Appendix). Only the published parameter estimates and their sampling variances together with a set of second-stage explanatory variables are necessary for the elaboration and estimation of the conceptually richer framework.

[3] The likelihood ratio test on the homogeneity of variance hypothesis is described in P. Hoel, pp. 225–28.

TABLE 2—HOSPITAL COST INCREASES[a]

| | $a_i$ = | $b_0$ | + $b_1 \log AC_0$ | + $b_2 \log S_0$ | + $b_3 \log U_0$ | + $b_4AT$ | + $b_5T$ | + $b_6P$ | + $b_7M$ | + $b_8 \log K_0$ | $R^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| (1) | $\sigma^2_{b_i}$ | $.398\times10^{-1}$ | $.406\times10^{-3}$ | $.266\times10^{-4}$ | $.539\times10^{-3}$ | — | — | — | — | — | — |
| | $\sigma^{2*}_{b_i}$ | $.433\times10^{-1}$ | $.686\times10^{-3}$ | $.448\times10^{-4}$ | $.652\times10^{-3}$ | — | — | — | — | — | — |
| | $b_i^{**}$ | .319 | *$-.505\times10^{-1}$* | *$.201\times10^{-1}$* | $.657\times10^{-2}$ | — | — | — | — | — | .397 |
| | $\sigma^{2**}_{b_i^{**}}$ | $.377\times10^{-3}$ | $.258\times10^{-3}$ | $.161\times10^{-4}$ | $.513\times10^{-3}$ | — | — | — | — | — | — |
| (2) | $\sigma^2_{b_i}$ | $.483\times10^{-1}$ | $.484\times10^{-3}$ | — | $.559\times10^{-3}$ | $.696\times10^{-4}$ | $.437\times10^{-4}$ | — | — | — | — |
| | $\sigma^{2*}_{b_i}$ | $.573\times10^{-1}$ | $.797\times10^{-3}$ | — | $.706\times10^{-3}$ | $.140\times10^{-3}$ | $.641\times10^{-4}$ | — | — | — | — |
| | $b_i^{**}$ | .348 | *$-.604\times10^{-1}$* | — | $.279\times10^{-1}$ | *$.223\times10^{-1}$* | *$.195\times10^{-1}$* | — | — | — | .289 |
| | $\sigma^{2**}_{b_i^{**}}$ | $.447\times10^{-1}$ | $.338\times10^{-3}$ | — | $.522\times10^{-3}$ | $.346\times10^{-4}$ | $.329\times10^{-4}$ | — | — | — | — |
| (3) | $\sigma^2_{b_i}$ | a | $.288\times10^{-3}$ | — | $.396\times10^{-3}$ | — | — | $.544\times10^{-4}$ | $.405\times10^{-4}$ | — | — |
| | $\sigma^{2*}_{b_i}$ | — | $.431\times10^{-3}$ | — | $.606\times10^{-3}$ | — | — | $.822\times10^{-4}$ | $.682\times10^{-4}$ | — | — |
| | $b_i^{**}$ | — | $.263\times10^{-1}$ | — | *$.421\times10^{-1}$* | — | — | $.125\times10^{-2}$ | $.908\times10^{-3}$ | — | .934 |
| | $\sigma^{2**}_{b_i^{**}}$ | — | $.215\times10^{-3}$ | — | $.291\times10^{-3}$ | — | — | $.375\times10^{-4}$ | $.248\times10^{-4}$ | — | — |
| (4) | $\sigma^2_{b_i}$ | $.528\times10^{-1}$ | $.539\times10^{-3}$ | $.586\times10^{-4}$ | $.548\times10^{-3}$ | $.143\times10^{-3}$ | $.631\times10^{-4}$ | $.693\times10^{-4}$ | $.339\times10^{-4}$ | — | — |
| | $\sigma^{2*}_{b_i}$ | $.648\times10^{-1}$ | $.856\times10^{-3}$ | $.704\times10^{-4}$ | $.684\times10^{-3}$ | $.217\times10^{-3}$ | $.868\times10^{-4}$ | $.116\times10^{-3}$ | $.603\times10^{-4}$ | — | — |
| | $b_i^{**}$ | .381 | *$-.557\times10^{-1}$* | *$.176\times10^{-1}$* | $.546\times10^{-2}$ | $.225\times10^{-2}$ | $.932\times10^{-2}$ | $-.388\times10^{-2}$ | $.536\times10^{-2}$ | — | .392 |
| | $\sigma^{2**}_{b_i^{**}}$ | $.495\times10^{-1}$ | $.388\times10^{-3}$ | $.507\times10^{-4}$ | $.549\times10^{-3}$ | $.104\times10^{-3}$ | $.493\times10^{-4}$ | $.446\times10^{-4}$ | $.207\times10^{-4}$ | — | — |
| (5) | $\sigma^2_{b_i}$ | .295 | $.530\times10^{-3}$ | $.587\times10^{-4}$ | $.620\times10^{-3}$ | $.140\times10^{-4}$ | $.623\times10^{-4}$ | $.695\times10^{-4}$ | $.342\times10^{-4}$ | $.144\times10^{-2}$ | — |
| | $\sigma^{2*}_{b_i}$ | .452 | $.836\times10^{-3}$ | $.686\times10^{-4}$ | $.778\times10^{-3}$ | $.213\times10^{-3}$ | $.865\times10^{-4}$ | $.118\times10^{-3}$ | $.624\times10^{-4}$ | $.246\times10^{-2}$ | — |
| | $b_i^{**}$ | *.986* | *$-.586\times10^{-1}$* | *$.204\times10^{-1}$* | $-.112\times10^{-1}$ | $.608\times10^{-3}$ | $.941\times10^{-2}$ | $-.272\times10^{-2}$ | $-.473\times10^{-2}$ | $-.423\times10^{-1}$ | .241 |
| | $\sigma^{2**}_{b_i^{**}}$ | .246 | $.388\times10^{-3}$ | $.540\times10^{-4}$ | $.692\times10^{-3}$ | $.104\times10^{-3}$ | $.486\times10^{-4}$ | $.447\times10^{-4}$ | $.206\times10^{-4}$ | $.961\times10^{-3}$ | — |

[a] In the event that the constant term in the unweighted regression proved insignificant, it was suppressed in the weighted regression. This was done in order to avoid possible spurious correlation. For those weighted regressions where this was not possible, the calculated coefficient of determination must be treated with care.

*Note:* $\hat{a}_i$ ≡ estimated first-stage parameter
$AC_0$ ≡ cost per patient day in the initial period
$S_0$ ≡ number of beds in the initial period
$U_0$ ≡ hospital occupancy rate
$AT$ ≡ dummy variable indicating hospital's advanced teaching status
$T$ ≡ dummy variable indicating hospital's teaching status
$P$ ≡ dummy variable indicating hospital's location in Pittsburgh
$M$ ≡ dummy variable indicating hospital's location in an urban area other than Pittsburgh
$K_0$ ≡ index of hospital's size relative to other hospitals in the initial period
$\sigma^2_{b_i}$ ≡ calculated sampling variance using classical formula in ordinary least squares case
$\sigma^{2*}_{b_i}$ ≡ calculated sampling variance using formula which allows for the presence of heteroscedasticity in ordinary least squares case
$b_i^{**}$ ≡ coefficient estimate using generalized least square methods
$\sigma^{2**}_{b_i^{**}}$ ≡ calculated sampling variance in generalized least squares case

Italicized coefficient significant at .05 level.

the new generalized least squares in Table 5 present a variety of significant explanatory variables.[4] The estimated size coefficients do not appear to be random numbers. Rather, variations in the coefficients are susceptible to systematic explanation. A change in the size of a hospital can significantly change average costs. Note for the time span considered, virtually all changes involved increases in hospital size. Thus the regressions in Table 5 suggest that increases in size will increase the average costs of already relatively large hospitals, hospitals with high initial average costs and hospitals in Pittsburgh, while reducing the average costs of advanced teaching hospitals and hospitals which, prior to expansion, have relatively high rates of utilization. While the results are quite plausible, especially when initial average costs and initial rates of utilization are taken as proxies for overall managerial competence, they should nonetheless be regarded with care. Only the coefficient on $\log U_0$ seems almost completely impervious to change in specification. Nonetheless, the

[4] When the Laves' pooled regressions are reestimated with the generalized least squares approach suggested here, allowing for heteroscedasticity, the size coefficient likewise becomes significant. For example, when (4−1) is reestimated we find

$$\log I = -\underset{(16.29)}{.458} \log J + \underset{(5.24)}{.122} \log K + \underset{(67.8)}{.056}t + \underset{(22.6)}{.067}F$$

TABLE 3—VARIATION IN UTILIZATION COEFFICIENT[a]

| $a_2$ = | $b_0$ | + $b_1 \log K_0$ | + $b_2 \log S_0$ | + $b_3 \log U_0$ | + $b_4 \log AC_0$ | + $b_5AT$ | + $b_6T$ | + $b_7P$ | + $b_8M$ | + $R^2$ |
|---|---|---|---|---|---|---|---|---|---|---|
| (1) $\sigma^2_{b_i}$ | .110x10$^{3}$ | .840 | — | — | — | — | — | — | — | — |
| $\sigma^{2*}_{b_i}$ | .265x10$^{3}$ | .202x10 | — | — | — | — | — | — | — | — |
| $b_i^{**}$ | *.264x10$^{2}$* | *−.234x10* | — | — | — | — | — | — | — | .280 |
| $\sigma^{2**}_{b_i^{**}}$ | .460x10$^{2}$ | .350 | — | — | — | — | — | — | — | — |
| (2) $\sigma^2_{b_i}$ | — | .947x10$^{-4}$ | — | — | — | — | — | .266x10$^{-1}$ | .213x10$^{-1}$ | — |
| $\sigma^{2*}_{b_i}$ | — | .286x10$^{-3}$ | — | — | — | — | — | .522x10$^{-1}$ | .479x10$^{-1}$ | — |
| $b_i^{**}$ | — | *.281x10$^{-1}$* | — | — | — | — | — | *−.265* | *−.381* | .384 |
| $\sigma^{2**}_{b_i^{**}}$ | — | .312x10$^{-4}$ | — | — | — | — | — | .181x10$^{-1}$ | .107x10$^{-1}$ | — |
| (3) $\sigma^2_{b_i}$ | — | .121x10$^{-3}$ | — | — | — | .458x10$^{-1}$ | .264x10$^{-1}$ | .373x10$^{-1}$ | .214x10$^{-1}$ | — |
| $\sigma^{2*}_{b_i}$ | — | .384x10$^{-3}$ | — | — | — | .107 | .427x10$^{-1}$ | .697x10$^{-1}$ | .468x10$^{-1}$ | — |
| $b_i^{**}$ | — | *−.304x10$^{-1}$* | — | — | — | .128 | .614x10$^{-1}$ | *−.335* | *−.382* | .611 |
| $\sigma^{2**}_{b_i^{**}}$ | — | .389x10$^{-4}$ | — | — | — | .197x10$^{-1}$ | .701x10$^{-1}$ | .256x10$^{-1}$ | .112x10$^{-1}$ | — |
| (4) $\sigma^2_{b_i}$ | — | — | — | .366x10$^{-3}$ | .119 | .462x10$^{-1}$ | .267x10$^{-1}$ | .376x10$^{-1}$ | .216x10$^{-1}$ | — |
| $\sigma^{2*}_{b_i}$ | — | — | — | .113x10$^{-2}$ | .325 | .106 | .425x10$^{-1}$ | .698x10$^{-1}$ | .468x10$^{-1}$ | — |
| $b_i^{**}$ | — | — | — | *−.517x10$^{-1}$* | −.547 | .127 | .604x10$^{-1}$ | *−.337* | *−.389* | .604 |
| $\sigma^{2**}_{b_i^{**}}$ | — | — | — | .119x10$^{-3}$ | .481x10$^{-1}$ | .201x10$^{-1}$ | .205x10$^{-1}$ | .261x10$^{-1}$ | .113x10$^{-1}$ | — |
| (5) $\sigma^2_{b_i}$ | — | .799x10$^{-1}$ | .377x10$^{-1}$ | .254 | .288 | .894x10$^{-1}$ | .395x10$^{-1}$ | .364x10$^{-1}$ | .214x10$^{-1}$ | — |
| $\sigma^{2*}_{b_i}$ | — | .253 | .689x10$^{-1}$ | .682 | .614 | .169 | .600x10$^{-1}$ | .774x10$^{-1}$ | .555x10$^{-1}$ | — |
| $b_i^{**}$ | — | *−.613* | −.155 | *.112x10* | .367x10$^{-1}$ | .228 | .750x10$^{-1}$ | *−.367* | *−.366* | .664 |
| $\sigma^{2**}_{b_i^{**}}$ | — | .284x10$^{-1}$ | .296x10$^{-1}$ | .116 | .179 | .606x10$^{-1}$ | .286x10$^{-1}$ | .231x10$^{-1}$ | .970x10$^{-2}$ | — |

[a] See note to Table 2.

results do confirm that in the case where estimated parameters are taken as dependent variables, the generalized least squares procedure proposed here (a procedure which may be regarded as a sensible reconciliation of the two methods proposed by the Laves) can lead to a significant and substantively important change in conclusions.

TABLE 4—VARIATIONS IN FIRST HALF COEFFICIENT[a]

| | $b_0$ | + $b_1 \log K_0$ | + $b_2 \log S_0$ | + $b_3 \log U_0$ | + $b_4 \log AC_0$ | + $b_5AT$ | + $b_6T$ | + $b_7P$ | + $b_8M$ | $R^2$ |
|---|---|---|---|---|---|---|---|---|---|---|
| (1) $\sigma^2_{b_i}$ | — | — | .107x10$^{-3}$ | .141x10$^{-2}$ | .104x10$^{-2}$ | — | — | — | — | — |
| $\sigma^{2*}_{b_i}$ | — | — | .169x10$^{-3}$ | .277x10$^{-2}$ | .212x10$^{-2}$ | — | — | — | — | — |
| $b_i^{**}$ | — | — | *.175x10$^{-1}$* | .959x10$^{-1}$ | *−.866x10$^{-1}$* | — | — | — | — | .712 |
| $\sigma^{2**}_{b_i^{**}}$ | — | — | .697x10$^{-4}$ | .781x10$^{-3}$ | .561x10$^{-3}$ | — | — | — | — | — |
| (2) $\sigma^2_{b_i}$ | — | — | — | .132x10$^{-2}$ | .965x10$^{-3}$ | — | — | .182x10$^{-3}$ | .135x10$^{-3}$ | — |
| $\sigma^{2*}_{b_i}$ | — | — | — | .296x10$^{-2}$ | .209x10$^{-2}$ | — | — | .323x10$^{-3}$ | .238x10$^{-3}$ | — |
| $b_i^{**}$ | — | — | — | *.539x10$^{-1}$* | −.358x10$^{-1}$ | — | — | *−.339x10$^{-1}$* | *−.522x10$^{-1}$* | .788 |
| $\sigma^{2**}_{b_i^{**}}$ | — | — | — | .718x10$^{-3}$ | .539x10$^{-3}$ | — | — | .109x10$^{-3}$ | .919x10$^{-4}$ | — |
| (3) $\sigma^2_{b_i}$ | — | — | — | .136x10$^{-2}$ | .989x10$^{-3}$ | .223x10$^{-3}$ | .151x10$^{-3}$ | — | — | — |
| $\sigma^{2*}_{b_i}$ | — | — | — | .271x10$^{-2}$ | .194x10$^{-2}$ | .161x10$^{-2}$ | .511x10$^{-3}$ | — | — | .720 |
| $b_i^{**}$ | — | — | — | *.108* | −.862x10$^{-1}$ | *.208x10$^{-1}$* | −.106x10$^{-1}$ | — | — | — |
| $\sigma^{2**}_{b_i^{**}}$ | — | — | — | .795x10$^{-3}$ | .591x10$^{-3}$ | .110x10$^{-3}$ | .155x10$^{-3}$ | — | — | — |
| (4) $\sigma^2_{b_i}$ | — | — | — | .124x10$^{-2}$ | .909x10$^{-3}$ | .272x10$^{-3}$ | .157x10$^{-3}$ | .229x10$^{-3}$ | .127x10$^{-3}$ | — |
| $\sigma^{2*}_{b_i}$ | — | — | — | .287x10$^{-2}$ | .201x10$^{-2}$ | .999x10$^{-3}$ | .247x10$^{-3}$ | .602x10$^{-3}$ | .223x10$^{-3}$ | — |
| $b_i^{**}$ | — | — | — | *.586x10$^{-1}$* | −.405x10$^{-1}$ | .169x10$^{-2}$ | .608x10$^{-2}$ | *−.343x10$^{-1}$* | *.483x10$^{-1}$* | .800 |
| $\sigma^{2**}_{b_i^{**}}$ | — | — | — | .694x10$^{-3}$ | .524x10$^{-3}$ | .874x10$^{-4}$ | .119x10$^{-3}$ | .110x10$^{-3}$ | .894x10$^{-4}$ | — |
| (5) $\sigma^2_{b_i}$ | — | .102x10$^{-2}$ | .235x10$^{-3}$ | .171x10$^{-2}$ | .190x10$^{-2}$ | .565x10$^{-3}$ | .248x10$^{-3}$ | .258x10$^{-3}$ | .133x10$^{-2}$ | — |
| $\sigma^{2*}_{b_i}$ | — | .147x10$^{-2}$ | .363x10$^{-3}$ | .350x10$^{-2}$ | .326x10$^{-2}$ | .160x10$^{-2}$ | .421x10$^{-3}$ | .613x10$^{-3}$ | .244x10$^{-3}$ | — |
| $b_i^{**}$ | — | .313x10$^{-1}$ | .108x10$^{-1}$ | .305x10$^{-1}$ | −.695x10$^{-1}$ | .655x10$^{-2}$ | −.108x10$^{-1}$ | *−.276x10$^{-1}$* | *−.477x10$^{-1}$* | .799 |
| $\sigma^{2**}_{b_i^{**}}$ | — | .748x10$^{-3}$ | .169x10$^{-3}$ | .169x10$^{-3}$ | .132x10$^{-2}$ | .282x10$^{-3}$ | .172x10$^{-3}$ | .142x10$^{-3}$ | .899x10$^{-4}$ | — |

[a] See note to Table 2.

TABLE 5—VARIATION IN SIZE COEFFICIENT

| $a_3 =$ | $b_0$ | $+ b_1 \log K_0$ | $+ b_2 \log S_0$ | $+ b_3 \log U_c$ | $+ b_4 \log AC_3 +$ | $b_5 AT +$ | $b_6 T +$ | $b_7 P +$ | $b_8 M$ | $R^2$ |
|---|---|---|---|---|---|---|---|---|---|---|
| (1) | — | — | .106 | −1.46 | 1.20 | — | — | — | — | .101 |
| | | | (.607) | (2.89) | (2.65) | | | | | |
| (2) | — | — | — | −1.41 | 1.23 | — | — | .220 | .359 | .151 |
| | | | | (2.55) | (2.55) | | | (1.09) | (1.52) | |
| (3) | — | .109 | — | −1.68 | −.158 | −.362 | .233 | .521 | −.029 | .171 |
| | | (1.60) | | (2.11) | (.187) | (1.64) | (.956) | (1.80) | (.145) | |
| (4) | — | — | .688 | −1.50 | .865 | −.965 | −.090 | .431 | .081 | .216 |
| | | | (2.51) | (2.23) | (1.56) | (2.92) | (.322) | (1.68) | (.413) | |
| (5) | — | .912 | .644 | −2.05 | .033 | −.951 | −.117 | .588 | .067 | .226 |
| | | (1.38) | (2.35) | (2.63) | (.041) | (2.90) | (.421) | (2.11) | (.348) | |

*Note:* Numbers in parentheses are $t$-statistics.

## APPENDIX

Let the Laves' first-stage and second-stage equations be given by

$$\text{(A1)} \quad y_j = X_j a_j + u_j, \qquad j = 1, 2, \ldots, 74$$

$$\text{(A2)} \quad a_j = Z_j b, \qquad j = 1, 2, \ldots, 74$$

where $a_j$ and $b$ are 4-component and (4x$G$)-component vectors of known parameters, $X_j$ is a 14x4 matrix of known constants, $Z_j$ is defined as diag $(z_j, z_j, \ldots, z_j)$ where $z_j$ is a $G$-component vector of known constants and $u_j$ is a 14-component vector of observable random variables with zero means and variance-covariance matrix given by

$$\text{(A3)} \quad \begin{aligned} E(u_j u_k) &= \lambda_{jj} I_{14} && \text{when } j = k \\ &= 0 && \text{when } j \neq k \end{aligned}$$

The term $\lambda_{jj}$ is defined as being scalar.

In what follows it will be assumed that $X$ and $Z$ are such that every matrix that must be inverted is indeed invertible.

Consider now the estimation of $a_j$ and $b$ assuming the $\lambda_{jj}$ are known. Given that the $\lambda_{jk}$, $k \neq j$ are equal to zero, best linear unbiased estimates of the $a_j$ will be simply given by

$$\text{(A4)} \quad \hat{a}_j = (X_j' X_j)^{-1} X_j' y_j, \qquad j = 1, 2, \ldots, 74$$

From (A2), (A3), and (A4) it follows that

$$\text{(A5)} \quad \hat{a}_j = Z_j b + \{X_j' [\lambda_{jj} I]^{-1} X_j\}^{-1} X_j' (\lambda_{jj} I)^{-1} u_j \qquad j = 1, 2, \ldots, 74$$

The term $\hat{b}$ can be defined as the best linear unbiased estimator for (A5) such that

$$\text{(A6)} \quad \hat{b} = \left\{\sum_{j=1}^{74} Z_j' [X_j' (\lambda_{jj} I_{14})^{-1} X_j] Z_j\right\}^{-1} \cdot \left\{\sum_{j=1}^{74} Z_j' X_j' [\lambda_{jj} I_{14}]^{-1} y_j\right\}$$

Another approach would be to combine (A1) and (A2) as

$$\text{(A7)} \quad y_j = X_j Z_j b + u_j \qquad j = 1, 2, \ldots, 74$$

It is apparent that the best linear unbiased estimator of $b$ will be given by $b^*$

$$\text{(A8)} \quad b^* = \left\{\sum_{j=1}^{74} Z_j' X_j' [\lambda_{jj} I_{14}]^{-1} X_j Z_j\right\}^{-1} \cdot \left\{\sum_{j=1}^{74} Z_j' X_j' [\lambda_{jj} I_{14}]^{-1} y_j\right\}$$

Equations (A6) and (A8) are identical

$$\begin{aligned} \hat{b} &= \left\{\sum_{j=1}^{74} Z_j' [X_j' (\lambda_{jj} I_{14})^{-1} X_j] Z_j\right\}^{-1} \cdot \left\{\sum_{j=1}^{74} Z_j' [X_j' (\lambda_{jj} I_{14})^{-1} X_j] \hat{a}_j\right\} \\ &= \left\{\sum_{j=1}^{74} Z_j' [X_j' (\lambda_{jj} I_{14})^{-1} X_j] Z_j\right\}^{-1} \cdot \left\{\sum_{j=1}^{74} Z_j' \left(\frac{1}{\lambda_{jj}}\right) X_j' X_j (X_j' X_j)^{-1} X_j y_j\right\} \\ &= \left\{\sum_{j=1}^{74} Z_j' [X_j' (\lambda_{jj} I_{14})^{-1} X_j] Z_j\right\}^{-1} \end{aligned}$$

$$\cdot \left\{ \sum_{j=1}^{74} Z_j' X_j' (\lambda_{jj} I_{14})^{-1} y_j \right\}$$
$$= b^*$$

In the present case where $\lambda_{jj}$ is unknown it can be estimated by

$$\text{(A9)} \quad \hat{\lambda}_{jj} = \frac{1}{14} (y_j - X_j \hat{a}_j)'(y_j - X_j \hat{a}_j)$$

Such estimates can be easily shown to be consistent (as sample size approaches infinity) under general conditions. An estimate of $b$ can be obtained by substituting $\hat{\lambda}_{jj}$ in (A6) and (A8). The estimate of $b$ obtained from such a substitution will have the same asymptotic distribution as $\hat{b}$ or $b^*$.

## REFERENCES

P. Hoel, *Introduction to Mathematical Statistics,* 3d ed., New York 1962.

J. R. Lave and L. B. Lave, "Hospital Cost Functions," *Amer. Econ. Rev.,* June 1970, *60,* 379–95.

J. G. Williamson, "Relative Price Change, Adjustment Dynamics and Productivity Growth," *Econ. Develop. Cult. Change,* July 1971, *19,* 507-27.

# The Foundations of Money Illusion in a Neoclassical Micro-Monetary Model: Comment

By ROBERT W. CLOWER AND JOHN G. RILEY*

Considering the attention that economists have devoted to monetary neutrality and related phenomena in recent years, it would be remarkable if, as Richard Dusansky and Peter Kalman (hereafter D-K) have recently suggested,[1] the microfoundations of money illusion were not now well established. In truth, D-K's suggestion appears to reflect a flaw in their understanding of the problem rather than a gap in received doctrine. Our purpose in this comment is to set the record straight by restating and qualifying D-K's results.

Modifying slightly the notation of D-K (pp. 115–16), we start by supposing that the demand functions of a typical household are defined as solutions, $\bar{x}(p, y, L)$, $\bar{m}(p, y, L)$ of the Patinkinesque utility maximization problem,

$$\text{(1)} \qquad \text{Max } u = f(x, m, p)$$

subject to $px+m-y-L=0$

If we suppose that the preference function $f$ satisfies *only* the usual continuity and convexity assumptions, then it is obvious (though D-K take two pages to show this) that the functions $\bar{x}$ and $\bar{m}$ *may* exhibit money illusion in the sense of Don Patinkin; i.e., in general, the functions $\bar{x}$ and $\bar{m}$ will *not* be homogeneous of degree zero and one, respectively, in the variables $p$, $y$, and $L$.[2] As is well known, however, a sufficient condition for absence of money illusion in the functions defined by (1) is that the preference function $f$ be homogeneous of degree zero in the variables $m$ and $p$, which is simply to say that, in Paul Samuelson's words, "... money is evaluated only in terms of the work which it has to do" (1947, p. 119).[3]

This condition can be weakened mathematically by requiring only that

$$\text{(2)} \qquad f(x, \lambda m, \lambda p) = \lambda^k f(x, m, p)$$

which is logically equivalent to the inelegant sufficiency conditions that D-K work out so laboriously in their paper (pp. 118–20).[4] But the "generality" thus achieved is spurious, since the conventional assumption that $f$ is homogeneous of degree zero already serves completely to characterize the class of illusion-free demand functions that are derivable from ordinal preference functions.[5] To see this, note that setting $\lambda=1/p_1$ in (2) and rearranging gives us

$$\text{(3)} \qquad f(x, m, p) = p_1^k f(x, m/p_1, p/p_1)$$

Clearly, all marginal rates of substitution, and hence all "real" demand functions, depend only on $f(x, m/p_1, p/p_1)$, which is homogeneous of degree zero in $m$ and $p$. Thus D-K's attempt to clarify the foundations of money illusion—which starts with an

* Professor of economics and assistant professor of economics, respectively, University of California, Los Angeles. We are grateful to Peter W. Howitt of the University of Western Ontario and to Roger Kormendi of UCLA for helpful suggestions.

[1] See D-K (1974), p. 115. See also D-K (1972, 1973).

[2] See Patinkin, pp. 22–23.

[3] See also Samuelson (1968), p. 8.

[4] The equivalence follows from the elementary fact (apparently overlooked by D-K) that a function is homogeneous of degree $k$ in some subset of variables if and only if its partial derivatives are homogeneous of the same degree with respect to variables excluded from, and of one lower degree with respect to variables included in the subset.

[5] Moreover, with the sufficiency condition expressed as in (2) above, it becomes clear that for all positive values of $k$ a proportional increase in prices and initial asset values increases the welfare of a household despite the fact that it chooses an identical bundle! While this implication is not subject to the usual market tests, one could pick households at random and ask each to make a bid on the option of receiving a proportional subsidy on the present value of assets and paying a tax at the same proportional rate on all purchases. D-K's generalization would be supported by nonzero bids on this option!

unwarranted criticism of Patinkin's classic account of the problem (compare D-K, p. 117, with Patinkin (1956), pp. 68–71, 313–14, and Patinkin (1965), pp. 78–79, 457–60) succeeds only in muddying water that was formerly clear.

A final remark is in order concerning D-K's discussion (pp. 120–22) of "testable implications" of illusion-free demand functions. Unlike the Hicksian symmetry conditions of conventional value theory, which follow directly from standard continuity assumptions, all of D-K's symmetry implications require strong restrictions on the form of the utility function—conditions to which D-K refer obliquely in footnotes but nowhere state explicitly in their paper. Modifying a theorem due to Kalman and Michael Intriligator, it can be shown that these restrictions reduce to the requirement that $f(x, m, p)$ be expressible in the semi-separable form[6]

$$f(x, m, p) = f_1(p)f_2(x, m) + f_3(p) \tag{4}$$

With this arbitrary and implausible requirement imposed as a maintained hypothesis, it would be disingenuous (to put it mildly) to suggest that empirical investigation of symmetry conditions might serve to establish the presence or absence of illusion-free demand functions.

[6] Kalman and Intriligator obtain the form

$$f(x, m, p) = g(x, m) + h(p) + \alpha(px + m)$$

However, to make this consistent with the D-K assumptions we must require either that $\alpha = 0$ or that $f$ be homogeneous of degree one in $m$ and $p$. Assuming the former allows the additional generality of (2).

## REFERENCES

**R. Dusansky and P. J. Kalman**, "The Foundations of Money Illusion in a Neoclassical Micro-Monetary Model," *Amer. Econ. Rev.*, Mar. 1974, *64*, 115–22; "Errata," *Amer. Econ. Rev.*, Dec. 1974, *64*, 1101.

——— **and** ———, "The Real Balance Effect and the Traditional Theory of Consumer Behavior: A Reconciliation," *J. Econ. Theory*, Dec. 1972, *5*, 336–47; "Erratum," *J. Econ. Theory*, Feb. 1973, *6*, 107.

**P. J. Kalman and M. D. Intriligator**, "Generalized Comparative Statics with Applications to Consumer Theory and Producer Theory," *Int. Econ. Rev.*, June 1973, *14*, 473–86.

**D. Patinkin**, *Money, Interest and Prices*, 1st, 2d eds., New York 1956, 1965.

**P. A. Samuelson**, *Foundations of Economic Analysis*, Cambridge, Mass. 1947.

———, "What Classical and Neoclassical Monetary Theory Really Was," *Can. J Econ.*, Feb. 1968, *1*, 1–15.

# The Foundations of Money Illusion in a Neoclassical Micro-Monetary Model: Comment

By C. Robert Wichers*

In an interesting and valuable article in this *Review*, Richard Dusansky and Peter Kalman (D-K) recently discussed the foundations of money illusion in a neoclassical micro-monetary model. Their micro-monetary model specifies a utility function depending on the quantities of all commodities and of money, as well as on all commodity prices. Maximization of the utility function subject to a conventional budget constraint leads to demand functions: Dusansky and Kalman's main result is the formulation of two conditions that are sufficient to render all demand functions free of money illusion.

In the present paper, D-K's result is extended in two ways. First, the sufficient conditions for illusion-free demand functions are weakened. This is done in Section I. Second, the validity of the D-K result is extended to supply functions as well. This is done in Section II. Two conclusions are thereby reached. These conclusions differentiate between money and commodities. In Section III, the origin of this differentiation is traced. It is found that money appears different from commodities (in the D-K model) only because it serves as the unit of account. Thus, money illusion is, in fact, numeraire illusion.

## I

The Dusansky-Kalman model requires that

$$u = f(x_1, \ldots, x_{n-1}, m; p_1, \ldots, p_{n-1}) = f(x, m; p)$$

be maximized, subject to the budget constraint

$$b = \sum_1^{n-1} p_i x_i + m$$

* Assistant professor of economics, State University of New York, Stony Brook. I am indebted to an anonymous referee for a number of helpful suggestions.

Here $x = \{x_1, \ldots, x_{n-1}\}$ is the vector of purchased commodity quantities; $p = \{p_1, \ldots, p_{n-1}\}$ is the (positive) vector of associated prices; $m$ represents money holdings at the end of the period; $b$ is the sum of money holdings at the beginning of the period and income during the period. (D-K use separate symbols for the two components of $b$. Since, however, only the sum of the components plays a role in what follows, and not the components individually, there is no harm in using $b$.) To avoid corner solutions, the $x_i$ and $m$ are taken to be positive. The utility function is assumed so that a constrained maximum exists, and so that the demand functions are unique in a neighborhood of the maximizing solution or solutions.

Let $(\bar{x}_1, \ldots, \bar{x}_{n-1}, \bar{m})$ maximize the utility function, subject to the budget constraint. Then these first-order conditions hold at $(\bar{x}_1, \ldots, \bar{x}_{n-1}, \bar{m})$:

$$\frac{\partial}{\partial x_i} f(x, m; p) = \lambda p_i \qquad i = 1, \ldots, n-1$$

$$\frac{\partial}{\partial m} f(x, m; p) = \lambda$$

$$b = \sum_1^{n-1} p_i x_i + m$$

Here $\lambda$ is the value assumed by the usual Lagrange multiplier. Solving the first-order conditions gives the locally unique demand functions

$$\bar{x}_i = \bar{x}_i(p, b) \qquad i = 1, \ldots, n-1$$
$$\bar{m} = \bar{m}(p, b)$$

By definition, the demand functions $\bar{x}_i(p, b)$ are illusion free if and only if they are homogeneous of degree zero in $p$ and $b$.

Dusansky and Kalman's main result is now:

(1) if all $\partial f(x, m; p)/\partial x_i$ are homogeneous of any degree $k$ in $m$ and $p$, and
(2) if $\partial f(x, m; p)/\partial m$ is homogeneous of degree $k-1$ in $m$ and $p$, then
(3) the demand functions are illusion free, and
(4) $\bar{m}(p, b)$ is homogeneous of degree 1 in $p$ and $b$.

The proof given by Dusansky and Kalman is ingenious, and far from easy. A slight artifice yields a simpler proof, which, moreover, leads immediately to weakened versions of conditions (1) and (2).

The artifice consists of introducing two $n$-vectors, $P=\{P_1, \ldots, P_n\}$ and $X=\{X_1, \ldots, X_n\}$, defined by

$$\{P_1, \ldots, P_n\} = \{p_1, \ldots, p_{n-1}, b\} \quad \text{and}$$
$$\{X_1, \ldots, X_n\} = \{x_1, \ldots, x_{n-1}, m/b\}$$

It is easy to see that

$$m = bX_n = P_nX_n \tag{5}$$

The utility function, $f(x, m; p)$, may now be written as $f(X_1, \ldots, X_{n-1}, P_nX_n; p)$, hence as $g(X_1, \ldots, X_{n-1}, X_n; p, P_n)$, that is, $g(X; P)$:

$$f(x, m; p) = g(X; P)$$

A useful and readily verified result is

$$f(x, am; ap) = g(X; aP) \qquad \text{(any scalar } a \neq 0) \tag{6}$$

The budget constraint

$$b = \sum_1^{n-1} p_ix_i + m \quad \text{becomes}$$

$$P_n = \sum_1^{n-1} P_iX_i + P_nX_n$$

or $$P_n = \sum_1^{n} P_iX_i$$

D-K's main result, (1)–(4), will now be reformulated in terms of $X$ and $P$, and then proved. After that, conditions (1) and (2) are weakened.

Condition (1) says that for $i=1, \ldots, n-1$, $\partial f(x, am; ap)/\partial x_i = a^k \partial f(x, m; p)/\partial x_i$. By (6), this is equivalent to $\partial g(X; aP)/\partial X_i = a^k\partial g(X; P)/\partial X_i$. Condition (2) says that $\partial f(x, am; ap)/\partial(am) = a^{k-1}\partial f(x, m; p)/\partial m$ or $\partial f(x, am; ap)/\partial m = a^k \partial f(x, am; p)/\partial m$. By (5) and (6), this may be written as $\partial g(X; aP)/\partial(P_nX_n) = a^k\partial g(X; P)/\partial(P_nX_n)$, and hence as $\partial g(X; aP)/\partial X_n = a^k \partial g(X; P)/\partial X_n$.

Conclusion (3), $\bar{x}_i(ap, ab) = \bar{x}_i(p, b)$, becomes $\bar{X}_i(aP) = \bar{X}_i(P)$, $i=1, \ldots, n-1$. Conclusion (4) says that $\bar{m}$, which by (5) equals $b\bar{X}_n$, is homogeneous of degree 1 in $p$ and $b$. Thus, $\bar{X}_n$ is homogeneous of degree zero in $p$ and $b$, or $\bar{X}_n(aP) = \bar{X}_n(P)$.

Combining results gives this restatement of (1)–(4): if $\partial g(X; aP)/\partial X_i = a^k\partial g(X; P)/\partial X_i$ for some $k$, all $a \neq 0$, and $i-1, \ldots, n$, then $\bar{X}_i(aP) = \bar{X}_i(P)$ for $i=1, \ldots, n$. For notational convenience, denote $\{\partial g/\partial X_1, \ldots, \partial g/\partial X_n\}$ by grad $g(X; P)$, and $\{\bar{X}_1(\ ), \ldots, \bar{X}_n(\ )\}$ by $\bar{X}(\ )$. The statement then becomes

(7) if grad $g(X; aP) = a^k$ grad $g(X; P)$ for some $k$ and all $a \neq 0$,
(8) then $\bar{X}(aP) = \bar{X}(P)$.

This is the D-K result, translated in terms of $g$, $X$, and $P$. Its proof is given in the next paragraph.

Since the vector $\bar{X}$ maximizes $g(X; P)$ subject to the budget constraint $P_n = \sum_1^n P_iX_i$, the elements of $\bar{X}$ satisfy the first-order conditions

$$\text{grad } g(X; P) = \lambda P \tag{9}$$

$$P_n = \sum_1^n P_iX_i \tag{10}$$

Equation (9) may be read as: the vectors grad $g(X; P)$ and $P$ are collinear. Let now $P$ be multiplied by $a \neq 0$. In view of (7), grad $g(X; P)$ is then multiplied by $a^k$. Since $a$ and $a^k$ are scalars, multiplication by $a$ preserves the collinearity of grad $g(X; P)$ and $P$, thus replacing (9) with an equivalent statement. Obviously, multiplication of $P$ by $a$ will replace (10), too, with an equivalent

statement. It follows that the solution $\bar{X}$ of (9), (10) is left invariant, proving (8). The proof of (7)–(8) and, hence, of (1)–(4) is thereby completed.

The proof requires less than D-K's, which is based on the assumption that the constrained utility maximum is regular, as well as on the assumption that the utility function is twice differentiable.

The preceding suggests a way of weakening (1) and (2), the sufficient conditions for illusion-free demand functions. Clearly, the crux of the proof is that multiplication of $P$ by $a$ preserves the collinearity of the vectors grad $g(X; P)$ and $P$. It would therefore be sufficient to require that multiplication of $P$ by $a$ turn grad $g(X; P)$ into a nonzero multiple of itself; it would even suffice if grad $g(X; P)$ became the sum of (i) such a multiple and (ii) a multiple of $P$.

Formalizing, we require—instead of (7)—that there exist two functions, $h_1(a)$ and $h_2(a)$, wholly arbitrary except for the condition that $h_1(a)\neq 0$ when $a\neq 0$, and such that

$$\text{(11)}\quad \text{grad } g(X; aP) = h_1(a)\cdot \text{grad } g(X; P) + h_2(a)\cdot P \qquad \text{for all } a \neq 0$$

Reverting to initial notation yields the sought for, weakened version of (1) and (2), and hence a generalization of the entire statement (1)–(4):

If, for some $h_1(a)$ and $h_2(a)$, with $h_1(a)\neq 0$ when $a\neq 0$, one has both

$$\text{(12)}\quad \frac{\partial}{\partial x_i} f(x, am; ap) = h_1(a)\frac{\partial}{\partial x_i} f(x, m; p) + h_2(a)p_i \qquad \text{for all } a \neq 0 \text{ and all } i, \quad \text{and}$$

$$\text{(13)}\quad \frac{\partial}{\partial(am)} f(x, am; ap) = \frac{1}{a}h_1(a)\frac{\partial}{\partial m} f(x, m; p) + \frac{1}{a}h_2(a)b \qquad \text{for all } a \neq 0$$

(14) then all demand functions are illusion free, and

(15) $\bar{m}(p, b)$ is homogeneous of degree 1 in its arguments.

Setting $h_1(a)=a^k$ and $h_2(a)=0$ gives back (1)–(4).

## II

Both the Dusansky-Kalman result and its generalization (12)–(15) refer to demand functions only. Both are further generalized in the present section: their validity is extended to supply functions as well.

It may be recalled that $b=P_n$ represents money holdings at the beginning of the period, plus the amount of money acquired during the period. Let now $b_i$, $i=1, \ldots, n-1$, be the quantity of good $i$ possessed at the beginning of the period, plus the quantity of good $i$ exogenously acquired—that is, acquired in ways other than by trading—during the period. Introduction of $b_1, \ldots, b_{n-1}$ does not affect the utility function, which is still

$$f(x, m; p) = g(X; P)$$

The budget constraint does change; it becomes

$$\sum_1^{n-1} p_i b_i + b = \sum_1^{n-1} p_i x_i + m$$

reflecting that the value of initial holdings, including transfers, equals the value of eventual holdings. Using $P=\{p_1, \ldots, p_{n-1}, b\}$ and $X=\{x_1, \ldots, x_{n-1}, m/b\}$, and defining $B=\{B_1, \ldots, B_n\}$ as $\{b_1, \ldots, b_{n-1}, 1\}$, one can write the budget constraint also as

$$\sum_1^n P_i B_i = \sum_1^n P_i X_i$$

or

$$\sum_1^n P_i(X_i - B_i) = 0$$

For $i=1, \ldots, n-1$, the difference $x_i - b_i = X_i - B_i$ represents, when evaluated at $\bar{x}_i=\bar{X}_i$, either the individual's demand for good $i$ during the period or the individual's supply of good $i$ during the period, depending on whether $\bar{x}_i - b_i$ is positive or negative. Clearly, the demand for or supply of good $i$ is homogeneous of degree zero in $p$ and $b$ if and only if $\bar{x}_i$ $(=\bar{X}_i)$ is invariant un-

der scalar multiplication of the vector $P=\{p_1, \ldots, p_{n-1}, b\}$.

For $i=n$, $\overline{X}_i-B_i$ is $\overline{X}_n-B_n=\overline{m}/b-1$. Multiplication by $P_n(=b)$ gives $P_n(\overline{X}_n-B_n)=\overline{m}-b$. This quantity is, if positive, the amount of money acquired—through trading—during the period; if negative, it is the amount of money spent. It is convenient to call $\overline{m}-b$ the individual's demand for or supply of money during the period. This demand or supply is homogeneous of degree one in $p$ and $b$ if and only if $\overline{X}_n$ is invariant under scalar multiplication of $P$, as is easy to see.

The purpose of this section, it is recalled, is to extend both the D-K result and its generalization (12)–(15) to supply. The extension sought can now be formulated (and will be proved immediately after formulation):

If, for some $h_1(a)$ and $h_2(a)$, with $h_1(a)\neq 0$ when $a\neq 0$,

$$\text{(16)}\quad \operatorname{grad} g(X; aP) = h_1(a) \operatorname{grad} g(X; P) + h_2(a)P \quad \text{for all } a \neq 0$$

then $\overline{X}(aP)=\overline{X}(P)$

As noted, $\overline{X}(aP)=\overline{X}(P)$ implies that all commodity demand and supply functions are homogeneous of degree zero in $P$, and that the demand or supply function for money is homogeneous of degree one in $P$. In initial notation, (16) reads: if both $\partial f(x, am; ap)/\partial x_i=h_1(a)\partial f(x, m; p)/\partial x_i+h_2(a)p_i$ for all $a\neq 0$ and all $i$ and $\partial f(x, am; ap)/\partial(am)=(1/a)h_1(a)\partial f(x, m; p)/\partial m+(1/a)h_2(a)b$ for all $a\neq 0$, then $\bar{x}_i-b_i$ is homogeneous of degree zero in $p$ and $b$ $(i=1, \ldots, n-1)$, and $\overline{m}-b$ is homogeneous of degree one in the same variables.

For a proof, consider that $\overline{X}$ maximizes utility $g(X; P)$ subject to the budget constraint. That is to say, $\overline{X}$ satisfies the first-order conditions

$$\text{(17)}\quad \operatorname{grad} g(X; P) = \lambda P$$

$$\text{(18)}\quad \sum_1^n P_i(X_i - B_i) = 0$$

Let now $P$ be multiplied by some scalar $a\neq 0$. Equation (18) is then replaced with an equivalent statement, as is obvious. Equation (17), asserting the collinearity of grad $g(X; P)$ and $P$, is also replaced with an equivalent statement, for, as the condition in (16) implies, grad $g(X; aP)$ and $aP$ are collinear. It follows that multiplication of $P$ by $a$ does not affect the solution $\overline{X}$, which completes the proof.

For an application of (16), let the utility function be classically specified, as $u=f(x_1, \ldots, x_{n-1})$. This utility function contains neither money nor prices, and is therefore homogeneous (of degree zero) in money and prices. Thus, the condition in (16) is necessarily satisfied. (Formal confirmation is obtained by setting $h_1(a)\equiv 1$ and $h_2(a)\equiv 0$.) It follows that the conclusion of (16) holds. In words: if the utility function is classical, all commodity demand functions as well as all commodity supply functions are homogeneous of degree zero in $p_1, \ldots, p_{n-1}$, and $b$; furthermore, the demand or supply function for money is homogeneous of degree one in the same variables. For commodity *demand* functions, this result is familiar. Also, if all money is spent—as one would expect when no utility is assigned to money—final holdings will be zero, $\overline{m}=0$. The demand or supply function for money, $\overline{m}-b$, is then $-b$, which is indeed homogeneous of degree one in $p_1, \ldots, p_{n-1}, b$.

## III

Money is traditionally held to be different from commodities, and the conclusion of (16)—commodity demand and supply functions homogeneous of degree zero, demand or supply function for money homogeneous of degree one—is in agreement with that point of view. What is unusual is that (16) gives *mathematical* expression to the difference between money and commodities. This implies that, somewhere among the assumptions and definitions, a *mathematical* distinction between money and commodities must have been introduced. The purpose of the present section is to find out where this distinction was introduced, and in which form.

Dusansky and Kalman, who comment on this matter, attribute the observed differ-

ence between money and commodities to the fact that, in their approach, "... $m$ is a nominal variable, while the $x_i$ are real variables" (p. 120, fn. 5). This cannot be the reason, however, because D-K do not express either the nominal nature of $m$ or the real nature of the $x_i$ in mathematical terms. Their distinction between nominal and real variables is therefore wholly interpretive, and cannot give rise to the observed, mathematical difference between money and commodities. That difference must have a mathematical origin.

As it turns out, money is different—in the present context, at least—only because it serves as the unit of account. This will be proved in two ways. The first proof shows that, if a *commodity* is chosen as the numeraire, the demand or supply function for that commodity is homogeneous of degree one, whereas all other demand and supply functions, *including the one for money*, are homogeneous of degree zero. The second proof shows that if some nongood, such as the guinea, is chosen as the numeraire, *all* demand and supply functions are homogeneous of degree zero. The two assertions to be proved are not meant to imply (or deny) that the selection of a numeraire is a formality. The sole purpose of the proofs is to trace the origin of the observed, mathematical difference between money and commodities.

To simplify the discussion, let $x_n$ and $b_n$ replace $m$ and $b$. The utility function, thus far $f(x_1, \ldots, x_{n-1}, m; p_1, \ldots, p_{n-1})$, then becomes $f(x_1, \ldots, x_n; p_1, \ldots, p_{n-1})$, and the budget constraint

$$\sum_1^{n-1} p_i b_i + b = \sum_1^{n-1} p_i x_i + m$$

becomes

$$\sum_1^{n-1} p_i b_i + b_n = \sum_1^{n-1} p_i x_i + x_n$$

This may also be written as

$$\sum_1^{n} p_i b_i = \sum_1^{n} p_i x_i$$

since $p_n \equiv 1$ anyway.

Let now some commodity be chosen as the numeraire—good 1, say. It is assumed that money does not thereby cease to exist; that is, it is assumed that a good may be called "money" even if some *other* good serves as the unit of account. The arguments of the utility function now include $p_n$, the price of money in terms of good 1, but they do not include $p_1$, since $p_1 \equiv 1$. Thus, the utility function becomes $f(x_1, \ldots, x_n, p_2, \ldots, p_n)$. The budget constraint,

$$\sum_2^{n} p_i b_i + b_1 = \sum_2^{n} p_i x_i + x_1$$

may be written as

$$\sum_1^{n} p_i b_i = \sum_1^{n} p_i x_i$$

using $p_1 \equiv 1$. Comparison with the preceding paragraph shows that the budget constraint is left unchanged (formally at least) when the role of numeraire is transferred from good $n$ to good 1; only the utility function is affected. The utility function now depends on $p_n$ but not on $p_1$, and before it depended on $p_1$ but not on $p_n$. Thus, when the two utility functions are maximized subject to the (identical) budget constraints, the implications differ only in that 1 and $n$ are interchanged. Specifically, the demand for or supply of good 1 is homogeneous of degree 1 in $p_2, \ldots, p_n$, and $b_1$; all other demand and supply functions—including the one for money—are homogeneous of degree zero in the same variables.

This completes the first proof that, in the present context, money differs from commodities only because it serves as the numeraire, as unit of account. Were it not for its role of numeraire, money would be—in the present context again—an ordinary commodity.[1]

For the second proof, choose the guinea as numeraire (in principle, any nongood would do). It will be shown that all demand and supply functions are homogeneous of degree

[1] It is of interest to compare this conclusion with one of Robert Clower's. According to Clower, money would be an ordinary commodity if it were not for its role of medium of exchange. The other two functions of money —store of value, unit of account—do not, in Clower's opinion, make money essentially different from commodities.

zero, implying, in particular, that money ceases to be different from (other) commodities once it loses its role of numeraire. The economy is taken to have $n$ goods, the $n$th of which is money. It is again assumed that a good may be called money even if it does not serve as the unit of account. For the sake of notational convenience, $x_n$ and $b_n$ will continue to replace $m$ and $b$.

All $n$ goods have prices, in terms of guineas, and none of these prices is definitionally equal to 1. That is to say, every price is a variable. The utility function therefore becomes $u=f(x_1, \ldots, x_n; p_1, \ldots, p_n)$, a function containing as many quantities of goods as it contains prices. The budget constraint

$$\sum_{1}^{n} p_i(x_i - b_i) = 0$$

appears the same as before. There is a difference, however: this time, none of the prices needs to be unity.

The problem of maximizing $f(x_1, \ldots, x_n; p_1, \ldots, p_n)$ subject to

$$\sum_{1}^{n} p_i(x_i - b_i) = 0$$

is formally the same as that of maximizing $g(X_1, \ldots, X_n; P_1, \ldots, P_n)$ subject to

$$\sum_{1}^{n} P_i(X_i - B_i) = 0$$

considered in the preceding section. Consequently, (16) holds, *mutatis mutandis*.

If, for some $h_1(a)$ and $h_2(a)$, with $h_1(a) \neq 0$ when $a \neq 0$, one has grad $f(x_1, \ldots, x_n; ap_1, \ldots, ap_n) = h_1(a)$ grad $f(x_1, \ldots, x_n; p_1, \ldots, p_n) + h_2(a)\{p_1, \ldots, p_n\}$ for all $a \neq 0$, then $\bar{x}_i(ap_1, \ldots, ap_n) = \bar{x}_i(p_1, \ldots, p_n)$. The conclusion implies that the $\bar{x}_i$, and hence the demand and supply functions $\bar{x}_i - b_i$ are homogeneous of degree zero in $p_1, \ldots, p_n$. This completes the second proof that, in the present context, money differs from commodities only in that money serves as the numeraire.

Considering that in the second proof both the utility function and the budget constraint are symmetric with respect to all goods, the symmetry of the just-proved result is not surprising. Were there such a thing as a demand or supply function for guineas, asymmetry would immediately return; for that function would be homogeneous of degree one in its arguments.

For an application of the just-proved result, let the numeraire be a nongood, like the guinea, and let the utility function be specified as

$$u = f(x_1, \ldots, x_n)$$

This utility function—which contains money, through the presence of $x_n$—does not depend on prices. As a consequence, the by now familiar condition for homogeneous demand and supply functions is satisfied. (Formal confirmation is obtained by setting $h_1(a) \equiv 1$ and $h_2(a) \equiv 0$.) It follows that all demand and supply functions, including the one for money, are homogeneous of degree zero in $p_1, \ldots, p_n$. Naturally, for this conclusion to hold it is essential that utility be assigned to money, and that no utility be assigned to the numeraire.

## IV. Summary and Conclusion

Sections I and II are summarized in (16). Section III proved in two different ways that the Dusansky-Kalman model distinguishes money from commodities solely by giving money the role of numeraire. Put differently, it is money's function of unit of account that sets money and commodities apart, in the D-K model. Money's function of store of value is reflected in the appearance of money in the utility function; in this, however, money is indistinguishable from other goods. Money's function of medium of exchange is not reflected in the model, is not translated in mathematical terms; in this, then, money is also indistinguishable from other goods.

## REFERENCES

R. W. Clower, "A Reconsideration of the Microfoundations of Monetary Theory," *Western Econ. J.*, Dec. 1967, *6*, 1–9.

R. Dusansky and P. J. Kalman, "The Foundations of Money Illusion in a Neoclassical Micro-Monetary Model," *Amer. Econ. Rev.*, Mar. 1974, *64*, 115–22.

# The Foundations of Money Illusion in a Neoclassical Micro-Monetary Model: Reply

By Richard Dusansky and Peter J. Kalman*

Robert Clower and John Riley's (hereafter, C-R) "analysis" of money illusion rests entirely on the following claims:

1) That their equation (2) is "logically equivalent" to our (hereafter, D-K) sufficiency conditions presented in the 1974 issue of this *Review*.
2) That the assumption of degree zero homogeneity (in the variables $m$ and $p$) of the utility function "serves completely to characterize the class of illusion-free demand functions" derivable from ordinal utility theory.
3) That it is possible to reinstate the usual Slutsky properties by adopting the "semi-separable" utility function given in their equation (4).

We will show that each of these claims is false.

The first two claims can be analyzed jointly by employing a simple counterexample. Consider the following utility function:

$$u(x, m, p) = g(x, m, p) + h(p) \tag{1}$$

where $g(x, m, p)$ is any real valued twice continuously differentiable function, strictly quasi concave and increasing in $(x, m)$ and homogeneous (of any degree $k$) in $m$ and $p$, and where $h(p)$ is any real valued twice continuously differentiable *non*-homogeneous function of all prices. This utility function, $u(x, m, p)$, is *not* homogeneous in both $m$ and $p$; in particular, it is not homogeneous of degree $k$ and it is not homogeneous of degree zero. Thus, it does *not* satisfy either of the C-R conditions in 1) and 2) above. C-R's remarks would suggest, if they were true, that the demand functions derivable from (1) *must* exhibit money illusion. But such is not the case. The resulting demand functions are indeed free of money illusion. To see this, examine the first-order conditions for the maximization of (1) subject to the budget constraint:

$$(2)\quad \frac{\dfrac{\partial u(x, m, p)}{\partial x_i}}{\dfrac{\partial u(x, m, p)}{\partial x_j}} = \frac{\dfrac{\partial g(x, m, p)}{\partial x_i}}{\dfrac{\partial g(x, m, p)}{\partial x_j}} = \frac{p_i}{p_j}$$

$$i = 1, \ldots, j-1, j+1, \ldots, N-1$$

$$\frac{\dfrac{\partial u(x, m, p)}{\partial m}}{\dfrac{\partial u(x, m, p)}{\partial x_j}} = \frac{\dfrac{\partial g(x, m, p)}{\partial m}}{\dfrac{\partial g(x, m, p)}{\partial x_j}} = \frac{1}{p_j}$$

$$\sum_{i=1}^{N-1} p_i x_i + m = Y + L$$

If all money prices and initial balances were to increase by some $\alpha$ (this is the usual money-illusion experiment), these first-order conditions would yield the unique solution $(x, \alpha m)$. Hence *the commodity demand functions are free of money illusion.* (This should not be surprising. The homogeneity property of $g(x, m, p)$ implies that $\partial g(x, m, p)/\partial x_i$ is homogeneous of degree $k$ in $m$ and $p$, and that $\partial g(x, m, p)/\partial m$ is homogeneous of degree $k-1$ in $m$ and $p$, so that the D-K (1974, p. 118) conditions for the absence of money illusion are satisfied.)

The counterexample in (1) thus clearly demonstrates that C-R are wrong in both of their first two assertions. Since the utility function in (1) satisfies the D-K sufficiency conditions while violating C-R's equation (2), the alternative sufficiency condition in their (2) *cannot* be "logically equivalent"; although (2) is violated, it is still possible (given the D-K conditions are satisfied) to

* State University of New York at Stony Brook and Harvard University, respectively.

have homogeneous of degree zero commodity demand functions. Furthermore, since the utility function in (1) satisfies the D-K conditions while violating the zero-homogeneity of utility function assumption, the latter *cannot* possibly serve to "completely characterize the class of illusion-free demand functions." At best it is an overly restrictive sufficiency condition. *Under the D-K sufficiency conditions it is possible to have demand behavior which is free of money illusion without imposing the restriction that the utility function be homogeneous of any degree.*[1] Consequently, C-R's equations (2), (3), and their associated discussion—which focus on homogeneous utility functions—are *irrelevant* to the D-K paper.[2]

We now turn to the third C-R claim: that the semi-separable utility function in their equation (4)

$$f(x, m, p) = e(p)\cdot g(x, m) + h(p)$$

in general satisfies the sufficient conditions for the reinstatement of the standard Slutsky properties (i.e., symmetry and negative semi-definiteness of the Slutsky matrix). The easiest way to see that this claim is false is to examine the Slutsky system for the type of model under discussion. This system is:[3]

$$(3)\quad \frac{\partial x_i}{\partial p_k} + x_k\frac{\partial x_i}{\partial y} = \lambda\frac{D_{ki}}{D} - \sum_{j=1}^{n-1}\frac{\partial^2 f}{\partial x_j\partial p_k}\frac{D_{ji}}{D} - \frac{\partial^2 f}{\partial m\partial p_k}\frac{D_{ni}}{D} = S_{ik},\qquad i, k = 1, \ldots, n-1$$

where $D$ is the last bordered Hessian and $D_{ij}$ is the cofactor of element $i, j$ in the Jacobian (see our 1974 paper, p. 116). The expression for $S_{ki}$ is given by:

$$(4)\quad \frac{\partial x_k}{\partial p_i} + x_i\frac{\partial x_k}{\partial y} = \frac{\lambda D_{ik}}{D} - \sum_{j=1}^{n-1}\frac{\partial^2 f}{\partial x_j\partial p_i}\frac{D_{jk}}{D} - \frac{\partial^2 f}{\partial m\partial p_i}\frac{D_{nk}}{D} = S_{ki},\qquad i, k = 1, \ldots, n-1$$

In order to have $S_{ik}=S_{ki}$, it is sufficient to establish the equality of the right-hand sides of (3) and (4). To do this, examine the C-R semi-separable utility function and compute the relevant second partial derivatives:

$$\frac{\partial^2 f}{\partial x_j\partial p_k} = \frac{\partial g(x, m)}{\partial x_j}\cdot\frac{\partial e(p)}{\partial p_k}$$

$$\frac{\partial^2 f}{\partial x_j\partial p_i} = \frac{\partial g(x, m)}{\partial x_j}\cdot\frac{\partial e(p)}{\partial p_i}$$

$$\frac{\partial^2 f}{\partial m\partial p_k} = \frac{\partial g(x, m)}{\partial m}\cdot\frac{\partial e(p)}{\partial p_k}$$

$$\frac{\partial^2 f}{\partial m\partial p_i} = \frac{\partial g(x, m)}{\partial m}\cdot\frac{\partial e(p)}{\partial p_i}$$

It is obvious that in general

$$\frac{\partial^2 f}{\partial x_j\partial p_k} \neq \frac{\partial^2 f}{\partial x_j\partial p_i},\qquad \frac{\partial^2 f}{\partial m\partial p_k} \neq \frac{\partial^2 f}{\partial m\partial p_i}$$

and that these second partial derivatives do not vanish. Similarly, in general $D_{ji}\neq D_{jk}$ and $D_{ni}\neq D_{nk}$. Hence the symmetry property is in general *not* reinstated.

The error in the C-R analysis of the symmetry property emanates from the process of pulling functional forms out of the air. This can be avoided by taking care to *derive* classes of functions which satisfy a specified separability condition, instead of guessing at functions that "might work." For the problem at hand, we now show how a general separable class of utility functions yielding the Slutsky properties can be derived.

We begin by introducing some convenient notation. Let the utility function $f(x, m, p)$ be denoted by $f(q, a)$ where $q=(x, m)$ and $a=(p, 1)$, let the constraint $y+L=\sum_1^{n-1} p_i x_i + m$ be denoted by $g(q, b)$ where $b=(a, L, y)$. The element in the $i$th row and $h$th column of the Slutsky matrix (see the authors, 1973, equations (31) and (32)) is:[4]

[1] Operationally this means that there is a larger admissible class of utility functions which is consistent with demand behavior free of money illusion.

[2] Similarly, the reservations expressed in C-R's fn. 5 do not apply to our paper. They apply only to C-R, who choose to focus on homogeneous utility functions.

[3] See the authors (1972), pp. 343–46 for derivation.

[4] Where $D_{hi}$ is the cofactor of the element in the $h$th row and the $i$th column of the Jacobian of the first-order conditions and $D$ is the determinant of the Jacobian.

$$S_{ih} = \lambda \frac{D_{hi}}{D} - \sum_{j=1}^{n} f_{q_j a_h} \frac{D_{ji}}{D}$$

or in matrix notation

$$(5) \qquad S = \frac{1}{D} [\lambda\Delta + F\Delta]$$

where
$$F = \nabla_q \nabla_a f$$
$$(\Delta)_{ji} = D_{ji}$$

So $S = S'$ if $F\Delta = \Delta F$. However, $F\Delta = \Delta F$ if

$$(6) \qquad F = I\phi(q, a)$$

where $I$ is the identity matrix. Hence we now need to construct a general solution to the system of partial differential equations stated in (6). By doing this, we obtain a general class of utility functions which yield symmetry of the Slutsky matrix. Rewrite (6) as

$$(7) \quad \frac{\partial^2 f}{\partial a_i \partial q_j} = \delta_{ij}\phi(q, a), \qquad i, j = 1, \ldots, n,$$

where $\delta_{ij} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases}$

Integrating with respect to $a_i$ we get

$$\frac{\partial f}{\partial q_j} = \mu(q, a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n)$$

$\forall i \neq j$, ($i$ taken one at a time, $i \in \{1, \ldots, n\}$). Hence we have

$$(8) \qquad \frac{\partial f}{\partial q_j} = \sigma(q, a_j)$$

Now taking partials of (8) with respect to $a_j$ we get

$$\frac{\partial^2 f}{\partial q_j \partial a_j} = \frac{\partial \sigma}{\partial a_j} = \beta(q, a_j)$$

$\forall j \in \{1, \ldots, n\}$. But by (7) $\beta \equiv \phi(q, a)\ \forall j$. Hence $\beta$ is independent of $j$. So we have

$$(9) \qquad \phi(q, a) = \phi(q)$$

Similarly for $q_j$ we get

$$(10) \qquad \phi(q, a) = \phi(a)$$

Hence from (9) and (10) we have $\phi(q, a) = K$ ($K$ is constant independent of $q$ and $a$). So in view of this, (7) becomes

$$(11) \qquad \frac{\partial^2 f}{\partial q_j \partial a_i} = \delta_{ij} K$$

Now for the case $i = j$ we have

$$\frac{\partial f}{\partial a_i} = Kq_i + k(q_1, \ldots, q_{i-1}, q_{i+1}, \ldots, q_n, a)$$

For the case $i \neq j$ we have (by (11))

$$\frac{\partial^2 f}{\partial q_j \partial a_i} = \frac{\partial k}{\partial q_j} = 0 \qquad \forall j \neq i$$

Hence $k = k(a)$. So we have

$$\frac{\partial f}{\partial a_i} = Kq_i + k(a)$$

Integrating with respect to $a_i$ we have

$$f = Ka_i q_i + \sigma(a) + \gamma(q, a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n)$$

Reversing the order of integration we obtain in a similar way

$$f = Ka_i q_i + \pi(q) + \rho(q_1, \ldots, q_{i-1}, q_{i+1}, \ldots, q_n, a)$$

Combining we obtain

$$f = Ka_i q_i + \pi(q) + \sigma(a) + \alpha(q_1, \ldots, q_{i-1}, q_{i+1}, \ldots, q_n, a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n) \qquad \forall i \in \{1, \ldots, n\}$$

Recalling (11), however, the most general form that

$$\alpha(q_1, \ldots, q_{i-1}, q_{i+1}, \ldots, q_n, a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n)$$

can take on is

$$\alpha = K \sum_{j=1}^{n} a_j q_j - Ka_i q_i$$

Thus we obtain the general solution to (7).

This is

$$(12)\quad f(q, a) = K \sum_{j=1}^{n} a_j q_j + \pi(q) + \sigma(a)$$

which is a general class of utility functions consistent with reinstatement of the symmetry property of the Slutsky system.[5]

[5] It can easily be shown that this utility function yields the negative semidefiniteness Slutsky property if $(K+\lambda)\geq 0$.

## REFERENCES


R. W. Clower and J. G. Riley, "The Foundations of Money Illusion in a Neoclassical Micro-Monetary Model: Comment," *Amer. Econ. Rev.*, Mar. 1976, *66*, 184–85.

R. Dusansky and P. J. Kalman, "The Real Balance Effect and the Traditional Theory of Consumer Behavior: A Reconciliation," *J. Econ. Theory*, Dec. 1972, *5*, 336–47.

——— and ———, "Erratum," *J. Econ. Theory*, Feb. 1973, *6*, 107.

——— and ———, "The Foundations of Money Illusion in a Neoclassical Micro-Monetary Model," *Amer. Econ. Rev.*, Mar. 1974, *64*, 115–22.

——— and ———, "Errata," *Amer. Econ. Rev.*, Dec. 1974, *64*, 1101.

# Faculty Salaries: Is There Discrimination by Sex, Race, and Discipline? Additional Evidence

By Emily P. Hoffman*

In a recent paper in this *Review*, Nancy Gordon, Thomas Morton, and Ina Braden (G-M-B) presented a model of faculty salary determination and results of an empirical test of their model. In this paper I present results of a replication of their study using data from another institution and results from tests of alternative models.

Faculty salary differentials can be explained in part by differences in individuals' characteristics. Jacob Mincer has stated that differences among individuals in their stock of human capital (for example, level of education, years of work experience) explains much of the variation in their earnings. College and university administrators claim that faculty salary depends on productivity in the areas of teaching, scholarship, and service. Differences in demand conditions among disciplines suggest including department or school among the factors thought to determine faculty salaries.[1]

There is no consensus among economists on a definition of labor market discrimination. G-M-B have an implicit definition of discrimination contained in their model. Their definition of labor market discrimination is that the sex differential in salary which is unexplained by differences in individual characteristics of age, seniority, education, rank, race, and discipline is due to sex discrimination. I propose a different definition of discrimination in that I omit rank as one of the relevant characteristics. Sex discrimination may occur through slower promotion rates for females than for males, in which case rank itself would reflect discrimination.[2]

I had access to a set of data from another institution, comparable to the G-M-B data. The two institutions are both large Ph.D. granting universities with similar faculty size. The variables in both studies are the same.

In Table 1, empirical results of a test of the G-M-B Model I using the new data set are presented alongside the G-M-B empirical results. Model I is as follows:

$$(1) \quad Ln\ W_i = a_0 + f(A_i) + a_5 L_i + d^n(E_i) + d^m(J_i) + d^q(S_i) + d^r(R_i) + d^t(D_i) + e_i$$

where, for the $i$th employee, $W_i$ represents salary, $A_i$ is age, $L_i$ is time at the university (seniority), $E_i$ is education, $J_i$ is academic rank, $S_i$ is sex, $R_i$ is race, $D_i$ is department, and $e_i$ is the disturbance term.[3] Only the coefficients for sex, race, seniority, and discipline are presented here.

There are few differences in results between the two sets of data. G-M-B find that women earn 9.5 percent less than men, while

* Assistant professor of economics, University of Missouri-Rolla. This paper was written while I was a doctoral candidate, University of Massachusetts. I wish to thank Ronald Ehrenberg and Ronald Oaxaca for their helpful suggestions and encouragement, and Richard Hoffman for his valuable computer programming assistance. Support by doctoral dissertation grant #DL 91-25-74-27 from the U.S. Department of Labor, Manpower Administration, Manpower Development Training Act is gratefully acknowledged.

[1] Another approach, used by Barbara Reagan and Betty Maynard, is to group departments by entrance salary levels. Their justification for this is that academic labor markets operate as internal labor markets; supply and demand forces are felt primarily at the entry level and have little effect on senior level faculty salaries. The internal labor market concept was developed by Peter Doeringer and Michael Piore.

[2] Janice Madden defines three types of sex discrimination: wage discrimination occurs when wage differentials are not based on productivity differences; occupational discrimination occurs when the number of females employed in each occupation is not based on productivity differences; cumulative discrimination occurs when a factor has low productivity due to past discrimination, so that the individual's characteristics incorporate discrimination. In this paper I demonstrate only the existence of wage discrimination.

[3] This is the notation of the G-M-B article, to which the reader is referred for a more complete explanation.

Table 1—Comparison of Selected Estimated Regression Coefficients of the G-M-B Model for G-M-B and Hoffman Data

| G-M-B Data | $\beta$ | $t$-value | Hoffman Data | $\beta$ | $t$-value |
|---|---|---|---|---|---|
| Sex, Female=1 | −.095 | 7.0 | Sex, Female=1 | −.074 | 7.1 |
| Race, Black=1 | .127 | 5.5 | Race, Black=1 | .074 | 4.1 |
| Seniority | −.005 | 5.4 | Seniority | −.002 | 2.5 |
| Discipline | | | Discipline | | |
| Fine Arts | .002 | 0.1 | Arts & Humanities | −.071 | 5.0 |
| Social Science | .096 | 4.3 | Social & Behavioral Science | −.036 | 2.3 |
| Quantitative Social Science | .186 | 7.5 | Natural Science & Math | −.026 | 1.7 |
| Engineering & Science | .139 | 7.3 | Engineering | .039 | 2.1 |
| Health | .188 | 9.7 | Public Health | −.051 | 2.2 |
| Law | .159 | 2.2 | Business Administration | .039 | 2.0 |
| Medicine | .402 | 16.2 | Physical Education | −.078 | 3.5 |
| Education | .189 | 9.0 | Graduate School Programs | −.002 | 0.1 |
| Other | Reference Group | | Natural Resources | −.062 | 3.7 |
| | | | Other | −.030 | 0.6 |
| | | | Education | Reference Group | |
| $\bar{R}^2$ | .794 | | $\bar{R}^2$ | .799 | |

I find sex causes a 7.4 percent difference in salary. Blacks earn significantly more than whites, which is probably due to a limited supply of blacks relative to an increased demand, as the administration attempted to comply with Affirmative Action legislation.

Model I was then tested with the rank variables omitted, using my data for males and females combined, and for each sex separately (with the sex variable omitted).[4] With this alternative definition of discrimination, the coefficient on sex more than doubles, from −.074 to −.162 ($t$=10.5).[5] The $R^2$ was .547 for the regression for males and females combined. For the female sample, age and seniority variables are more significant when rank is omitted. For both the male sample and the all faculty sample, the coefficient for seniority is no longer significant once rank is omitted. Negative coefficients for age beyond age 35 suggests young faculty members' salaries rise faster than those of older faculty, which is consistent with the internal labor market concept. At this university, there was a large increase in demand for faculty over the last decade, which had a greater effect on entry level faculty salaries than on those of senior faculty.

The age, seniority, and degree variables were then replaced by the following two variables: 1) the quadratic form of years of potential experience measured as years since highest degree, which is an estimate of years of professional experience; and 2) years of schooling since high school. It was then possible to estimate the coefficients of the human capital postschooling investment model.

The human capital postschooling investment model performs well in predicting faculty salaries.[6] The $R^2$ was .595 for the regression fitted for males and females combined. The estimated coefficients can be interpreted as follows: each year of experience increases one's salary at most 4.3 percent per year, with diminishing returns occurring; and each

[4] In each case when male and female models were separately fitted, the estimated coefficients in each model were found to differ significantly.

[5] In a study of salary differentials of professional employees of a firm, Burton Malkiel and Judith Malkiel found that discrimination against females occurred through lower job level assignments to females than to males with the same characteristics. They measured discrimination both with and without job level as an explanatory variable, and found greater discrimination when job level was omitted.

[6] Using 1970 National Science Foundation Register data for Ph.D.s and the human capital approach, George Johnson and Frank Stafford found that female faculty earn less than male faculty. They attribute part of the male-female salary differential to discrimination, and part to differences between males and females in human capital investments, i.e., less labor force participation and less on-the-job training for females.

year of schooling after high school increases one's salary an average of 7.6 percent.[7] In this model, females earn 14.1 percent less than males ($t=9.8$).

One can compare the amount of discrimination measured as a residual predicted by each model under two alternative assumptions.[8] Under Assumption 1, I predict what the average male and female salaries would be if both were compensated by the male salary structure, while Assumption 2 is that males and females would be compensated by the female salary structure. Any resulting difference in male and female average salaries which is not explained by differences in average characteristics, when the same salary structure is used for both groups, is attributed to discrimination. The G-M-B model attributes between 29 and 35 percent of the salary differential by sex to discrimination, based on Assumptions 1 and 2, respectively. Omitting rank from their model and using my data, from 56 to 68 percent of the salary differential by sex is seen to be due to discrimination, based on Assumptions 1 and 2, respectively (for males and females with equal characteristics). Again we observe that omitting rank as an explanatory variable increases the measured amount of discrimination, almost doubling it.

The Gordon-Morton-Braden model gives an estimate of sex discrimination which is only half that given by my alternative model, which accounts for the possibility that discrimination against female faculty may occur through slower promotion rates. The human capital model is shown to perform well in explaining faculty salaries.

[7] This interpretation is subject to Sherwin Rosen's caveat concerning the difficulty of identifying rate of return to schooling from a reduced form equation.

[8] This technique of measuring discrimination was developed by Ronald Oaxaca.

REFERENCES

P. Doeringer and M. Piore, *Internal Labor Markets and Manpower Analysis*, New York 1971.

N. Gordon, T. Morton, and I. Braden, "Faculty Salaries: Is There Discrimination by Sex, Race, and Discipline?," *Amer. Econ. Rev.*, June 1974, *64*, 419–27.

G. Johnson and F. Stafford, "The Earnings and Promotion of Women Faculty," *Amer. Econ. Rev.*, Dec. 1974, *64*, 888–903.

J. Madden, *The Economics of Discrimination*, Lexington 1973.

B. Malkiel and J. Malkiel, "Male-Female Pay Differentials in Professional Employment," *Amer. Econ. Rev.*, Sept. 1973, *63*, 693–705.

J. Mincer, "The Distribution of Labor Incomes: A Survey with Special Reference to the Human Capital Approach," *J. Econ. Lit.*, Mar. 1970, *8*, 1–26.

R. Oaxaca, "Male-Female Wage Differentials in Urban Labor Markets," *Int. Econ. Rev.*, Oct. 1973, *14*, 693–709.

B. Reagan and B. Maynard, "Sex Discrimination in Universities: An Approach through Internal Labor Market Analysis," *AAUP Bull.*, Spring 1974, *60*, 11–21.

S. Rosen, "Human Capital and the Internal Rate of Return," in G. Somers, ed., *Proc. 26th Annual Winter Meeting*, Ind. Rel. Res. Assn. Series, Madison 1974, 243–50.

# Default Risk, Homemade Leverage, and the Modigliani-Miller Theorem: Note

*By* KÅRE P. HAGEN*

In the March 1974 issue of this *Review* David Baron uses a stochastic dominance argument in an attempt to prove that the Modigliani-Miller (M-M) theorem is generally valid even in case of default risk. For this purpose he uses the familiar two-firm paradigm where both firms have identical probability distributions for gross returns, and one firm has some debt in its capital structure whereas the other firm is financed entirely by equity capital. If all equity investors in the levered firm also hold bonds in that firm or if all investors can borrow at the same nominal interest rate as firms, then he shows that in equilibrium both firms must have the same total market value.

Although there is nothing formally wrong in his arguments, Baron does not prove what he sets out to prove. What he proves seems to be the (fairly obvious) fact that at an equilibrium in a perfect capital market two firms with identical probability distributions for gross returns must have the same total market value. This does not imply, however, that the equilibrium value of a firm is independent of its capital structure. In fact, the common market value of the two firms will in general be dependent on the debt-equity ratio in the levered firm. To show this a simple counterexample will suffice.

## I

The conventional one-period model where investors invest at the beginning of the period while returns materialize at the end is used with the following notations:

- $X_j(\theta)$ = gross returns in firm $j$ if state of the world $\theta$ obtains, $\theta \in \Omega$, where $\Omega$ is the state space
- $R_j(\theta)$ = returns to equity in firm $j$ in state $\theta$
- $B_j(\theta)$ = returns to bondholders in firm $j$ in state $\theta$
- $D_j$ = total debt liability in firm $j$ (principal plus interest payments due at the end of the period)
- $\alpha_{ij}$ = the fraction investor $i$ holds of the shares of firm $j$
- $\beta_{ij}$ = the fraction investor $i$ holds of the bonds issued by firm $j$
- $S_j$ = market value of firm $j$'s equity
- $B_j$ = market value of firm $j$'s bonds
- $V_j \equiv S_j + B_j$ = total market value of firm $j$
- $Y_i(\theta)$ = investor $i$'s final wealth in state $\theta$
- $W_i$ = investor $i$'s initial wealth
- $\pi_i(\theta)$ = investor $i$'s probability assessments
- $U_i(Y_i(\theta))$ = investor $i$'s utility function (strictly concave)

Clearly,

$$R_j(\theta) = \text{Max}\,[0,\, X_j(\theta) - D_j]$$
$$B_j(\theta) = \text{Min}\,[D_j,\, X_j(\theta)]$$
$$\text{and } X_j(\theta) = R_j(\theta) + B_j(\theta) \qquad \text{for all } \theta \in \Omega$$

Two firms with identical gross return patterns over states of the world are assumed, i.e., $X(\theta) = X_j(\theta)$ for all $\theta \in \Omega$, $j = 1, 2$. Firm 1 is financed entirely by equity capital such that $V_1 \equiv S_1$, while the total debt liability in firm 2 is $D_2$. Short selling of stocks and bonds is generally possible. If $\beta_{ij} < 0$, investor $i$ is issuing bonds with the same return characteristics as those issued by firm $j$. Moreover, riskless lending and borrowing are available to the investors at a zero interest rate. Each investor owns initially a fraction $\delta_i$ of each firm and investors are assumed to rank portfolios according to expected utility of final wealth.

$$\text{Define } \hat{\Omega}(D_2) \equiv \{\theta \in \Omega \mid X(\theta) \geqq D_2\}$$
$$\text{and } \bar{\Omega}(D_2) \equiv \Omega \backslash \hat{\Omega}(D_2)$$

* Assistant professor of economics, The Norwegian School of Economics and Business Administration.

i.e., $\bar{\Omega}(D_2)$ are the states of the world in which firm 2 goes bankrupt.

Using the budget constraint, final wealth can be written as (1a) if $\theta \in \hat{\Omega}$ or (1b) if $\theta \in \bar{\Omega}$.

$$\text{(1a)} \quad Y_i(\theta) = W_i + \alpha_{i1}(X(\theta) - V_1) + \alpha_{i2}(X(\theta) - S_2) - \beta_{i2}B_2 + (\beta_{i2} - \alpha_{i2})D_2$$

$$\text{(1b)} \quad Y_i(\theta) = W_i + \alpha_{i1}(X(\theta) - V_1) + \beta_{i2}(X(\theta) - B_2) - \alpha_{i2}S_2$$

where $W_i = \delta_i(V_1 + V_2)$.

The first-order conditions for optimal portfolios for all $i$ can be written as:

$$\text{(2)} \qquad V_1 = \sum_{\theta \in \Omega} \omega_i(\theta) X(\theta)$$

$$\text{(3)} \qquad S_2 = \sum_{\theta \in \hat{\Omega}(D_2)} \omega_i(\theta)(X(\theta) - D_2)$$

$$\text{(4)} \qquad B_2 = \sum_{\theta \in \hat{\Omega}(D_2)} \omega_i(\theta) D_2 + \sum_{\theta \in \bar{\Omega}(D_2)} \omega_i(\theta) X(\theta)$$

$$\text{where } \omega_i(\theta) \equiv \frac{\pi_i(\theta) U_i'(Y_i(\theta))}{\sum_{\theta \in \Omega} \pi_i(\theta) U_i'(Y_i(\theta))}$$

The conditions (2), (3), and (4) together with the market-clearing conditions $\Sigma_i \alpha_{ij} = 1$, $\Sigma_i \beta_{ij} = 1$ for all $j$, characterize an equilibrium in the capital market.

The term $\omega_i(\theta)$ is investor $i$'s marginal rate of substitution between a unit return in state $\theta$ and a unit of riskless return. Hence, $\omega_i(\theta)$ can be interpreted as the implicit price in terms of the riskless asset investor $i$ is willing to pay for a claim to a unit return contingent on state $\theta$. As the price of a unit of riskless return is $\Sigma_{\theta \in \Omega}\, \omega_i(\theta) = 1$, the riskless asset is a numeraire.

As can be seen from (2), (3), and (4), $V_2 \equiv S_2 + B_2 = V_1$ so that the two firms will in equilibrium always have the same market value. But this common market value will in general depend on the debt level in the levered firm unless $\omega_i(\theta)$ is independent of $D_2$ for all $\theta$ and for all $i$. Hence we have the immediate result that the market value of the two firms is independent of the leverage in the levered firm if and only if $\omega_i(\theta)$ does not depend on $D_2$ for all $\theta$ and for all $i$.

If the market structure were complete, i.e., there are as many securities with linearly independent return patterns as states of the world, the return pattern on the optimal portfolios would be independent of the return structure of the individual securities in which case $\omega_i(\theta)$ would clearly be independent of $D_2$. Moreover, irrespective of the completeness of the market structure, if $\bar{\Omega}(D_2) = \phi$ for any relevant $D_2$ (no default risk), then $B_2 = D_2$ from (4) so that $D_2$ drops out of (1) which again would imply $\omega_i(\theta)$ to be independent of $D_2$.

In general, with incomplete markets and default risk in the levered firm, $\omega_i(\theta)$ will be independent of $D_2$ if and only if $\beta_{i2} = \alpha_{i2}$ for all $i$ as can be seen from (1). Hence, in an economy with incomplete markets for risk, a sufficient condition for firm values to be dependent on leverage is a nonzero default risk and that investors in equilibrium hold different fractions of the equity and bonds of the levered firm ($\alpha_{i2} \neq \beta_{i2}$).

## II

A numerical example may be in order to illustrate the above point. Assume two investors each of whom owns initially 50 percent of each firm, i.e., $W_1 = W_2 = 0.5V_1 + 0.5V_2$. The utility functions are assumed to be given by $U_1(Y_1) = Y_1 - Y_1^2/730$ and $U_2(Y_2) = Y_2 - Y_2^2/870$. First it is assumed that both firms are financed entirely by equity capital so that $V_j \equiv S_j$, $j = 1, 2$. Riskless lending and borrowing at a zero interest rate are available for both investors. There are three different states of the world labelled 1, 2, and 3. It is postulated that the following data are given:

| $\theta$ | 1 | 2 | 3 |
|---|---|---|---|
| $X_1(\theta) = X_2(\theta)$ | 150 | 200 | 300 |
| $\pi_1(\theta)$ | 1/2 | 1/4 | 1/4 |
| $\pi_2(\theta)$ | 1/4 | 1/4 | 1/2 |

It is fairly easy to verify that the equilibrium solution is given by $\alpha_{11} = \alpha_{12} = 1/3$, $\alpha_{21} = \alpha_{22} = 2/3$, $V_1 = V_2 = 193.8$.

The above example is now modified such that the firm labelled 2 is financed both by equity and debt. In all other respects the relevant data are the same as above. It is

assumed that $D_2=160$. Clearly $D_2/B_2$ will be the gross nominal interest rate on firm 2's bonds. Investors are assumed to be able to issue bonds on the same terms as firms. That means that investors can issue bonds with the same return characteristics as those of firm 2 at the same nominal interest rate. A personal bond issue takes place if $\beta_{i2}<0$. Moreover, riskless lending and borrowing at a zero interest rate are still available.

The total returns to shareholders and bondholders of the two firms will now be as follows:

| $\theta$ | 1 | 2 | 3 |
|---|---|---|---|
| $X_1(\theta)=R_1(\theta)$ | 150 | 200 | 300 |
| $R_2(\theta)$ | 0 | 40 | 140 |
| $B_2(\theta)$ | 150 | 160 | 160 |
| $X_2(\theta)$ | 150 | 200 | 300 |

As can be seen from the table, firm 2 goes bankrupt in state 1 so that there is a positive probability for bankruptcy for both investors. It is assumed again that each investor holds initially 50 percent of each firm, that is, 50 percent of the shares of firm 1 and 50 percent of the shares and bonds of firm 2.

After some tedious calculations, it is found that an equilibrium is characterized by

$$\begin{aligned}
V_1 &\equiv S_1 = 195 \\
S_2 &= 40 \\
B_2 &= 155 \\
V_2 &\equiv S_2 + B_2 = 195 \\
\alpha_{11} &= -6 - \beta_{12} \\
\alpha_{12} &= (20 + 3\beta_{12})/3 \\
\beta_{12} &= \text{arbitrary} \\
\alpha_{21} &= 7 + \beta_{12} \\
\alpha_{22} &= -(17 + 3\beta_{12})/3 \\
\beta_{22} &= 1 - \beta_{12}
\end{aligned}$$

As can be seen, the optimal portfolios are not unique. This is due to the fact that there are four securities with different return patterns (including the riskless opportunity) and just three states of the world. Only three of these return patterns are linearly independent. Hence, the investors can obtain their most preferred pattern of final wealth over states of the world by combining these four securities in an infinite number of ways. Note, however, that for any choice of $\beta_{12}$, $\alpha_{i2}\neq\beta_{i2}$, $i=1, 2$. For example, for $\beta_{12}=-6$, the solution is:

$$\begin{aligned}
\alpha_{11} &= 0 & \alpha_{21} &= 1 \\
\alpha_{12} &= 2/3 & \alpha_{22} &= 1/3 \\
\beta_{12} &= -6 & \beta_{22} &= 7
\end{aligned}$$

The nominal interest rate on risky bonds, $(D_2/B_2-1)100$, is 100/31 percent. Investor 1 issues six times the amount of risky bonds issued by firm 2 with the same return characteristics at the same nominal interest rate. The individual transactions corresponding to the above portfolios are summarized below:

| | Shareholdings Firm 1 | Shareholdings Firm 2 | Borrowing (−)/ Lending (+) Riskless | Borrowing (−)/ Lending (+) Risky |
|---|---|---|---|---|
| Investor 1 | 0 | 80/3 | 3295/3 | −930 |
| Investor 2 | 195 | 40/3 | −3295/3 | 1085 |
| Total | 195 | 40 | 0 | 155 |

The notable point in the above example is that although $V_1=V_2=195$, this common value is different from the case where both firms were financed by equity capital only.[1]

## III

Consider in general a one-period exchange economy with $s$ states of the world and $n$ firms with gross returns denoted by $X_j(\theta)$, $\theta\in\Omega$. Investor $i$'s income pattern over states of the world at the end of the period is given by

$$Y_i(\theta) = \sum_{j=1}^{n} \alpha_{ij}R_j(\theta) + \sum_{j=1}^{n} \beta_{ij}B_j(\theta) + rm_i, \quad \theta\in\Omega \tag{5}$$

where $m_i$ denotes riskless lending or borrowing and $r$ is one plus the riskless rate of interest. If $\beta_{ij}<0$, investor $i$ is issuing risky bonds

[1] In fact, if the levered firm in the above example borrows at least as much as to make the default risk positive, the values of the two firms will be independent of the particular debt level in the levered firm because we then have as many linearly independent securities as states of the world (complete markets).

with the same return characteristics as those issued by firm $j$. An investor's income pattern over states of the world can be represented by an $s$-component vector and it is assumed that each investor has a unique preference ordering over such income vectors. A feasible income pattern obtainable in the security markets is given by (5). The set of all obtainable income vectors will be called the feasible income space, the structure of which will depend on the generating return vectors $R_j(\theta)$ and $B_j(\theta)$ which in turn depend on the capital structures in the various firms. Let $\hat{Y}_i(\theta)$ denote the return pattern on investor $i$'s optimal portfolio in some feasible income space corresponding to some set of capital structures in the various firms. It is fairly obvious that firm $k$'s equilibrium value will be independent of its capital structure if and only if the income patterns $\hat{Y}_i(\theta)$ remain optimal for all $i$ in any feasible income space corresponding to any set of financial arrangements in firm $k$.[2] This holds if and only if at least one of the following two conditions is satisfied:[3]

(i) The feasible income space must remain unchanged for different debt-equity ratios in firm $k$.

(ii) The return patterns on the equilibrium portfolios corresponding to all the feasible income spaces which can be generated by varying the debt-equity ratio in firm $k$ must be contained in the intersection of all the feasible income spaces so created.

Condition (i) is always satisfied if there is no default risk in firm $k$ for all relevant debt levels (the original M-M theorem). If there is default risk in firm $k$, condition (i) holds if:

(ia) There are as many securities with linearly independent return patterns as there are states of the world. In that case the return vectors span the whole $s$-dimensional income space so that the income patterns $\hat{Y}_i(\theta)$ will remain optimal for any particular capital structure in firm $k$.

(ib) The return patterns $R_k(\theta)$ and $B_k(\theta)$ of firm $k$ are linearly dependent on the rest of the market for any debt level in firm $k$.

A special case of (ib) would be if any return pattern $B_k(\theta)$ which the firm can create could have been created by an investor.[4] He can do that by a limited liability arrangement where he buys equal shares of total stocks and bonds of firm $k$ and finances a varying part of these purchases by a personal loan pledging his shares and bonds in firm $k$ as the sole collateral. If the returns on his shares and bonds in firm $k$ are insufficient to meet the nominal payments due on the loan, the lenders get the returns on the pledged securities and the investor pays nothing but forfeits his securities in firm $k$. By varying the margin, i.e., his homemade leverage, the investor can replicate the return pattern on any conceivable bond issue in firm $k$.

If condition (i) does not hold, the firm will in fact create a new security with a unique return pattern by changing its capital structure. In that case condition (ii) must hold if the firm's value is to remain unchanged with different debt levels. It is easy to see that the intersection of the feasible income spaces so created is given by vectors of the form

$$Y_i(\theta) = rm_i + \sum_{j\neq k} \alpha_{ij}R_j(\theta) + \sum_{j\neq k} \beta_{ij}B_j(\theta) + \gamma_{ik}X_k(\theta)$$

That is, for any debt level in firm $k$ it must be optimal for each investor to hold the same fraction of firm $k$'s equity and bonds ($\alpha_{ik} = \beta_{ik} \equiv \gamma_{ik}$ for all $i$). For the M-M theorem to

[2] Sufficiency follows from the fact that if $\hat{Y}_i(\theta)$ are invariant to the capital structure in firm $k$, implicit prices $\omega_i(\theta)$ will be independent of the capital structure in that firm. If the value of firm $k$ shall remain the same for any capital structure in that firm, it is clearly also necessary.

[3] Condition (i) is obvious. Even though condition (i) does not hold, condition (ii) guarantees that the return patterns on the optimal portfolios corresponding to the various feasible income spaces must remain the same because they are obtainable in each of the incomes spaces so created. For example, let $\bar{S}$ and $\hat{S}$ be two feasible income spaces corresponding to two different debt levels in firm $k$ and let $\bar{Y}_i(\theta)$ and $\hat{Y}_i(\theta)$ be the return patterns on the optimal portfolios corresponding to these two situations. Since $\bar{Y}_i(\theta)\in\bar{S}\cap\hat{S}$ and $\hat{Y}_i(\theta)\in\bar{S}\cap\hat{S}$, we must have that $\hat{Y}_i(\theta)=\bar{Y}_i(\theta)$ for all $i$, provided that the optimal return patterns are unique which will always be the case with strictly concave utility functions.

[4] This is equivalent to Joseph Stiglitz's assumption, p. 788.

be globally true, i.e., to be valid for all $j$, the return patterns on the equilibrium portfolios satisfying condition (ii) must be of the form

$$Y_i(\theta) = rm_i + \sum_j \gamma_{ij} X_j(\theta)$$

where $\gamma_{ij}$ is the optimal fraction investor $i$ holds of the total stocks and bonds of firm $j$. That is, each investor must in equilibrium hold the same fraction of the total amount of shares and risky bonds issued by any firm. If investors are ranking portfolios according to expected utility, it is well known[5] that each investor will in equilibrium hold the same share of *all* the risky securities if and only if separation obtains in the individual portfolios and investors have the same expectations. That is, the return patterns on the equilibrium portfolios will be of the form $Y_i(\theta)=rm_i+\gamma_i\sum_j X_j(\theta)$ for all $i$. Hence, portfolio separation and homogeneous beliefs are clearly sufficient conditions for the M-M theorem to be valid in case of default risk.

## IV

Let us now reexamine Baron's proof of the M-M theorem under default risk in the light of the above arguments. It is not sufficient for the theorem to be valid that all equity investors in the levered firm also hold bonds in that firm, unless it would be optimal for each investor to hold the same fraction of the equity and bonds in that firm (condition (ii)). Whether his second assumption is sufficient depends on how one understands the assumption that investors are able to borrow at the same nominal interest rates as firms. On page 178 it seems that he means that an investor can replicate on his personal loan the same return pattern as on the bonds issued by the levered firm such that lenders would be indifferent between lending to the investor and buying the risky bonds, i.e., allowing $\beta_{ij}$ to be negative. This is, however, not sufficient for the theorem to be valid. The investors must be able to create any conceivable return pattern which the firm could have created itself by issuing different amounts of risky bonds, and for the lenders it must be immaterial whether these return patterns are created by individuals or by firms. By changing its capital structure the firm would in this case create an investment opportunity which is already available in the market and hence the feasible income space and firms' market values would remain unchanged. Regardless of the capital structure in firm $k$, the income patterns $\hat{Y}_i(\theta)$ will in this case remain optimal and can be obtained by buying shares and bonds in firm $k$ on margin and pledging these securities as the sole collateral. Hence, investors can undo the allocation effects from changing the debt-equity ratio in firm $k$ by a suitable adjustment of the homemade leverages on their personal loans. Those homemade leverages may, however, not coincide with the observed leverage in firm $k$.

If the above conditions are not satisfied, firms' equilibrium values will depend on their capital structures. It will, however, always be true that firms with identical gross returns have the same total market value. But this does not prove that the market values are independent of the way firms are financed.

[5] See David Cass and Stiglitz for a rigorous discussion of portfolio separation.

## REFERENCES

**D. P. Baron**, "Default Risk, Homemade Leverage, and the Modigliani-Miller Theorem," *Amer. Econ. Rev.*, Mar. 1974, *64*, 176–82.

**D. Cass and J. E. Stiglitz**, "The Structure of Investor Preferences and Asset Returns, and Separability in Portfolio Allocation," *J. Econ. Theory*, June 1970, *2*, 122–60.

**J. E. Stiglitz**, "A Re-Examination of the Modigliani-Miller Theorem," *Amer. Econ. Rev.*, Dec. 1969, *59*, 784–93.

# Default Risk and the Modigliani-Miller Theorem: A Synthesis

*By* DAVID P. BARON*

Several recent studies have extended the original results of Franco Modigliani and Merton Miller (M-M) to the case in which there is default risk on the debt obligations of a firm. These extensions are concerned with the relative values of firms in a risk class, with the absolute value of those firms, and with the preferences of investors for changes in capital structure. I have given conditions under which the relative values of firms in the same risk class are independent of the capital structure of any firm in that class, but this does not imply that the common value of those firms in the risk class will in general be independent of the capital structure as indicated by Joseph Stiglitz (Section III) and Kåre Hagen. If investors may create any return opportunities that a firm is able to create by altering its capital structure, the value of the firm will, however, be independent of its capital structure. For example, investors will be able to create the equivalent return opportunities by borrowing on personal account and pledging shares of an unlevered firm as collateral, since the investor can then borrow under the same conditions and at the same nominal interest rate as can the firm as indicated in my earlier paper. These results pertain to firm valuation, but investors are interested in the return on their portfolio and not directly in the value of firms. Vernon Smith has investigated the preferences of investors for the debt-equity ratio of a variable-scale firm and has indicated that investors are not indifferent to changes in that ratio and that investors may not prefer a change in the debt-equity ratio of a firm even if the value of the firm is increased by that change.[1]

In order to place these results in perspective, it is useful to return to the original M-M theorem that was given for firms with a fixed scale of investment. That theorem actually has three facets that may be stated as follows:

*I:* In the absence of default risk the values of all firms in the same risk class are equal regardless of their capital structures.[2]

*II:* In the absence of default risk the common value of firms in a risk class is independent of the capital structure of any firm.[3]

*III:* In the absence of default risk investor preferences are in agreement with changes in the value of firms in the sense that investors will prefer a change

* Professor of managerial economics and decision sciences, Graduate School of Management, Northwestern University. I wish to thank Mark Satterthwaite and Robert Taggart for their comments on an earlier version of this paper.

[1] As will be indicated below the differences among these results stem from differing assumptions made regarding the completeness of the market in which the firms' securities are traded. If investors may borrow on personal account and create the same return opportunities as firms create by altering their capital structure, the common value of the firms in a risk class will be independent of capital structure, and investors will be indifferent to changes in the capital structure. In my earlier paper, I gave a homemade leverage argument that implies that the capital market is complete in this sense. Without these homemade leverage opportunities or an equivalent completeness condition, the values of firms in a risk class and investor preferences will depend on a firm's financial structure.

[2] The proof of M-M's Proposition I is a proof of the first facet. In the introduction to their Proposition I, M-M state that they will "exhibit the mechanism determining the relative prices" (p. 268), and at the end of that proof, state that "we can conclude that in equilibrium we must have $V_2=V_1, \ldots$," (p. 271), where $V_1$ is the market value of an unlevered firm and $V_2$ is the market value of a levered firm. These results may be stated in an analogous manner for firms whose patterns of return differ by a multiplicative constant.

[3] This facet is stated by M-M in Proposition I: "that is, the market value of any firm is independent of its capital structure . . . " (p. 268). This facet is not directly proved by M-M, but it clearly holds.

in the capital structure of a firm if and only if the value of the firm is increased.[4] It follows from *II* that investors are indifferent to changes in the capital structure.

Stiglitz (Sections I and II) demonstrated that the second and third facets hold without the risk class assumption, so that the value of a firm is independent of the capital structure of any firm and all investors are indifferent to changes in capital structure. Default risk, however, can eliminate the equivalence among these three facets, so that while the implication of *I* continues to hold under the conditions I have given, the implications of *II* and *III* are no longer valid in general. Furthermore, shareholders may not prefer a change in capital structure of a firm even if that change increases the value of the securities they hold.

Given the divergence of these three facets, the question of the proper objective for financial management must be raised. The appropriate criterion is that a firm should undertake investments and capital structure changes when they are in the interests of shareholders. Under certain conditions guaranteed by a complete capital market, all shareholders can be shown to be unanimous with respect to their preferences for investment-financing decisions and to prefer an increase in investment (financed by debt) if and only if it increases the value of the equity of the firm. When the capital market is not complete, the investment-financing decisions of a firm can create new patterns of returns in the capital market, and investors will then revise their portfolios resulting in a new capital market equilibrium. Unanimity will then not occur in general, and the criterion of acting in the shareholders' interests will not provide an unambiguous guide for investment-financing decisions.

To investigate these issues, a model that includes those of M-M, Stiglitz, Smith, Hagen, and the author as special cases will be considered. That model is introduced in the next section in which the first facet of the M-M theorem is considered. The second and third facets are considered in Section II where conditions are given under which shareholder preferences will be in accord with changes in the value of the securities held by shareholders. Conclusions are offered in the final section.

## I. Relative Values and Homemade Leverage

The earnings $X_i(\theta)$ of firm $i$, $i=1, 2$, in a risk class are assumed to be given by[5]

$$X_i(\theta) = (1+\theta)k(M_i), \qquad k(0) = 0,$$
$$k'(M_i) \geq 0, \qquad M_i \geq 0$$

where $M_i$ is the investment or scale of the firm and $\theta$ is the net rate of return per unit of $k(M_i)$. Each firm eventually has nonincreasing returns to scale so that $k''(M_i) \leq 0$ for $M_i \geq \overline{M} > 0$. Firm 1 is assumed to be financed only by equity with its value denoted by $V_1$, while the value $V_2$ of firm 2 equals the value $S_2$ of its equity plus the debt $D_2$ that it issues. Firm 2 is assumed to begin with no outstanding debt and then to issue debt $D_2$ at a (gross) interest rate $R_2$ established endogenously in the capital market.

The model will be considered under two assumptions regarding the scale of the firm. First, the scale will be taken to be fixed as considered by M-M, Stiglitz, Hagen, and the author, so that increases in debt represent a purely financial transaction of selling debt and distributing the proceeds to shareholders. In the second case, the initial scale $M^*$ of the firms will be assumed to be fixed with increases in scale financed by debt, so that $M_2 = M^* + D_2$.

An investor $i$ is taken to have an initial endowment consisting of an $\bar{\alpha}_{i1}$ share of the equity of the unlevered firm, an $\bar{\alpha}_{i2}$ share of the equity of the levered firm, and savings (borrowing if negative) of $\bar{m}_i$ that yields the gross risk-free interest rate of one. For the case in which the debt proceeds are distributed to shareholders, the investor chooses

[4] The third facet is not directly stated by M-M but is implied by facet *II* and by their criterion that investments and financing should be judged by their effect on the market value of the firm.

[5] Similar results can be obtained with a more general return function but this formulation is used to correspond with Smith's model.

a new portfolio composition ($\alpha_{i1}$, $\alpha_{i2}$, and $\beta_{i2}$) subject to the budget constraint

$$(1)\quad \bar{m}_i + \bar{\alpha}_{i1}V_1 + \bar{\alpha}_{i2}S_2 = m_i + \alpha_{i1}V_1 + \alpha_{i2}S_2 + (\beta_{i2} - \bar{\alpha}_{i2})D_2$$

where $\beta_{i2}$ is the proportion of the debt of the levered firm purchased. For the case in which the debt is used to finance investment, the budget constraint is that in (1) without the term $(-\bar{\alpha}_{i2}D_2)$.[6] The portfolio return $Y_i(\theta)$ for the investor is (2a) if $-1 \leq \theta \leq \theta^*$ and (2b) if $\theta \geq \theta^*$.[7]

$$(2a)\quad Y_i(\theta) = m_i + \alpha_{i1}(1+\theta)k(M_1) + \beta_{i2}(1+\theta)k(M_2)$$

$$(2b)\quad Y_i(\theta) = m_i + \alpha_{i1}(1+\theta)k(M_1) + \beta_{i2}R_2D_2 + \alpha_{i2}((1+\theta)k(M_2) - R_2D_2)$$

where $\theta^* = R_2D_2/k(M_2) - 1$ is the level of $\theta$ below which the firm defaults on its debt obligations. An investor is assumed to maximize his expected utility $E_iU_i(Y_i(\theta))$ of portfolio return, where $E_i$ denotes expectation with respect to his assessed density function $f_i(\theta)$ and $U_i$ is a strictly concave, measurable utility function. The securities market will be assumed to be competitive in that investors believe that their portfolio decisions do not affect the market values of the securities. Using the budget constraint to eliminate $m_i$, the necessary conditions for an optimal portfolio are

$$(3)\quad \int_{-1}^{\infty} \omega_i(\theta)(1+\theta)k(M_1)d\theta - V_1 = 0$$

$$(4)\quad \int_{\theta^*}^{\infty} \omega_i(\theta)((1+\theta)k(M_2) - R_2D_2)d\theta - S_2 = 0$$

$$(5)\quad \int_{-1}^{\theta^*} \omega_i(\theta)(1+\theta)k(M_2)d\theta + \int_{\theta^*}^{\infty} \omega_i(\theta)R_2D_2d\theta - D_2 = 0$$

where

$$\omega_i(\theta) = U_i'(Y_i(\theta))f_i(\theta) \Big/ \int_{-1}^{\infty} U_i'(Y_i(\theta))f_i(\theta)d\theta$$

is the implicit price of a unit of portfolio return if $\theta$ obtains. Since

$$\int_{-1}^{\infty} \omega_i(\theta)d\theta = 1$$

$\omega_i(\theta)$ represents the marginal rate of substitution between a certain return of one dollar on the risk-free asset and the return $Y_i(\theta)$ when $\theta$ occurs.

These necessary conditions will be attainable if 1) investors may sell short or 2) the optimal portfolio contains positive holdings of the securities of each firm. In these cases a capital market equilibrium is characterized by (3), (4), and (5)[8] for each investor and by the market-clearing conditions $\sum_i \hat{\alpha}_{i1} = 1$, $\sum_i \hat{\alpha}_{i2} = 1$, and $\sum_i \hat{\beta}_{i2} = 1$, where $(\hat{\ })$ denotes optimal. The existence of such an equilibrium is assumed here. Adding (4) and (5) and comparing the resulting expression with (3) indicates that for $M_2 = M_1$, the value of the unlevered firm equals the value of the levered firm.[9] The first facet of the M-M theorem is thus valid in general if investors may sell short or hold positive shares of the securities. In terms of the second facet, if the levered firm may affect its market value by a purely financial transaction, the value of the unlevered firm must change by an equal amount.

Whether or not a purely financial transaction will affect the value of the firms depends on the nature of the capital market. The capital market will be said to be "financially complete" if any pattern of return that a fixed-scale firm may create by a purely

[6] The model studied by Smith will be referred to as a paid-in capital model in that initial holdings $\bar{\alpha}_{i1}$ and $\bar{\alpha}_{i2}$ equal zero and the scale $M_2$ of the firm is determined by the amount $S_2+D_2$ paid in by investors. This model will be considered in Section III.

[7] The net return $\theta$ will be assumed to be greater than or equal to minus one.

[8] This presumes that $Y_i(\theta)$ is positive in the optimal portfolio for all $\theta$ or that if $Y_i(\theta)$ is negative for some $\theta$, the investor can cover the negative return with other assets. Smith considered the case in which $Y_i(\theta)$ negative implies personal bankruptcy, in which case the investor is not able to borrow at the risk-free interest rate.

[9] If the scales of the two firms are different, $M_1 \neq M_2$, then $V_1/k(M_1) = V_2/k(M_2)$.

financial transaction already exists in the market or can be created from existing securities. In that case a change in capital structure will not alter the space of returns available to the investor, and the implicit prices will thus be independent of the capital structure of a fixed-scale firm. The capital market will be financially complete under a variety of assumptions including a complete set of contingent claims markets, a complete set of Arrow-Debreu securities, or a spanning set of securities.[10] Another condition under which the capital market will be financially complete is if investors may create homemade leverage by borrowing using shares of the unlevered firm as collateral.[11] For fixed-scale firms, homemade leverage allows the investor to duplicate the specific pattern of returns generated by the securities of the levered firm, and thus the values of the firms will be equal as indicated in my earlier paper. In the absence of market imperfections the investor may also use homemade leverage to create any pattern of returns that the levered firm is able to create. The capital market is then financially complete, and as will be shown in the next section, both the second and third facets of the M-M theorem hold for fixed-scale firms. If the change in capital structure introduces a "new security" (a new pattern of returns) into the capital market so that the capital market is financially incomplete, the second and third facets will not in general be in accord.

For the case in which the scale of the firm may be changed, the capital market will be said to be "scale and financially complete" if returns identical to those that can be generated by varying the scale of a firm and its financing exist separately in the capital market, can be formed from other securities, or can be created through homemade leverage. For example, if a firm possesses stochastic constant returns to scale ($k'(M)M/k(M)=1$), changes in scale will not alter the set of available returns if there is no default risk. Nonconstant returns to scale may, however, alter the set of available returns if the capital market is not scale and financially complete. The next section considers the relationship between firm valuation and investor preferences in complete and incomplete capital markets.

## II. Firm Valuation and Investor Preferences

### A. *Firm Valuation*

The effect of debt financing on the value of a firm can be investigated by totally differentiating the first-order conditions (4) and (5) to obtain, respectively,

$$(6)\quad dS_2=\int_{\theta^*}^{\infty}\omega_i(\theta)((1+\theta)k'(M_2)dM_2 - (R_2dD_2+D_2dR_2))d\theta + \int_{\theta^*}^{\infty}d\omega_i(\theta)((1+\theta)k(M_2)-R_2D_2)d\theta$$

$$(7)\quad dD_2=\int_{-1}^{\theta^*}\omega_i(\theta)(1+\theta)k'(M_2)dM_2d\theta + \int_{\theta^*}^{\infty}\omega_i(\theta)(R_2dD_2+D_2dR_2)d\theta + \int_{-1}^{\theta^*}d\omega_i(\theta)(1+\theta)k(M_2)d\theta + \int_{\theta^*}^{\infty}d\omega_i(\theta)R_2D_2d\theta$$

where $d\omega_i(\theta)$ is the change in the implicit

[10] The implicit prices may also be independent of the debt level under certain other conditions in which separation results as indicated by Hagen and Frank Milne (1974). The relationship between an Arrow-Debreu model and other models that yield the same type of results, such as those with spanning sets, is developed by Milne (1974), Roy Radner, Steinar Ekern (1974, 1975), Ekern and Robert Wilson, Peter Diamond, and David Cass and Joseph Stiglitz.

[11] This is equivalent to borrowing at a stochastic interest rate $R_1(\theta, Z)$ defined by

$$R_1(\theta,Z)=\begin{cases}(1+\theta)k(M_1)/Z & \text{if } -1\le\theta\le\theta^0(Z)\\ R_1(Z) & \text{if } \theta^0(Z)\le\theta\end{cases}$$

where $R_1(Z)$ is the nominal interest rate on the borrowing, $\theta^0(Z)=R_1(Z)Z/k(M_1)-1$ is the default level for such borrowing, and $Z$ is a size parameter. If $Z=D_2$, then $R_1(Z)$ equals the interest rate on the debt of the levered firm. The investor may thus borrow at the same interest rate as can the firm, as indicated in my earlier paper. Smith, however, has shown that the investor is not able to borrow at the same interest rate as the firm if the investor may pledge only the shares of the levered firm as collateral.

price. Adding the expressions in (6) and (7) yields

$$(8)\quad dV_2 = \int_{-1}^{\infty} \omega_i(\theta)(1+\theta)k'(M_2)dM_2 d\theta + \int_{-1}^{\infty} d\omega_i(\theta)(1+\theta)k(M_2)d\theta$$

For the case of a fixed scale ($dM_2=0$), a purely financial transaction leaves the value of the firm unchanged if the capital market is financially complete (since $d\omega_i(\theta)=0$ for all $\theta$).[12] The values of both firms will thus not be affected by a change in the financial structure of the levered firm, since $V_1 = V_2$.

When the scale is not fixed but the securities market is scale and financially complete, the change in the value of the firm resulting from debt-financed investment may be determined by adding (4) and (5), solving for $\int_{-1}^{\infty} \omega_i(\theta)(1+\theta)d\theta$, and substituting into (8) to obtain

$$(9)\quad dV_2 = k'(M_2)V_2 dM_2/k(M_2)$$

The total value of the firm will thus increase (decrease) as the result of debt-financed investment if the marginal product $k'(M_2)$ is positive (negative).

### B. *Investor Preferences in a Complete Market*

Investors, however, are not directly interested in the value of the firm but rather are concerned with the expected utility resulting from their portfolio return. An investor will prefer a change in the debt level of a firm if and only if that change yields an increase in expected utility. The resulting (normalized) change in expected utility is given by

$$(10)\quad \frac{dE_iU_i}{E_iU_i'} = \int_{-1}^{\theta^*} \omega_i(\theta)dY_i(\theta)d\theta + \int_{\theta^*}^{\infty} \omega_i(\theta)dY_i(\theta)d\theta$$

where

$$dY_i(\theta) = \hat{\alpha}_{i1}(1+\theta)k'(M_1)dM_1 + \hat{\beta}_{i2}(1+\theta)k'(M_2)dM_2 + d\hat{m}_i \quad \text{if } -1 \leq \theta < \theta^*$$

and

$$dY_i(\theta) = \hat{\alpha}_{i1}(1+\theta)k'(M_1)dM_1 + \hat{\alpha}_{i2}((1+\theta)k'(M_2)dM_2 - (R_2dD_2 + D_2dR_2)) + \hat{\beta}_{i2}(R_2dD_2 + D_2dR_2) + d\hat{m}_i \quad \text{if } \theta^* \leq \theta$$

and $\quad E_iU_i' = \int_{-1}^{\infty} U_i'(Y_i(\theta))f_i(\theta)d\theta$

The change in expected utility can be related to the change in the value of the equity of the firm by substituting (6) and (7) into (10) to obtain[13]

$$(11)\quad \frac{dE_iU_i}{E_iU_i'} = d\hat{m}_i + \hat{\alpha}_{i2}dS_2 + \hat{\beta}_{i2}dD_2 + \hat{\alpha}_{i2}\int_{\theta^*}^{\infty} d\omega_i(\theta) \cdot ((1+\theta)k(M_2) - R_2D_2)d\theta + \hat{\beta}_{i2}\int_{-1}^{\theta^*} d\omega_i(\theta)(1+\theta)k(M_2)d\theta + \hat{\beta}_{i2}\int_{\theta^*}^{\infty} d\omega_i(\theta)R_2D_2 d\theta$$

For a purely financial transaction in which the funds raised by issuing debt are distributed to shareholders, the change in investor borrowing (lending) is

$$(12)\quad d\hat{m}_i = (\bar{\alpha}_{i1} - \hat{\alpha}_{i1})dV_1 + (\bar{\alpha}_{i2} - \hat{\alpha}_{i2})dS_2 + (\bar{\alpha}_{i2} - \hat{\beta}_{i2})dD_2$$

With a financially complete capital market and a purely financial transaction, the expression in (11) reduces to

$$(13)\quad \frac{dE_iU_i}{E_iU_i'} = \bar{\alpha}_{i2}dS_2 + \bar{\alpha}_{i2}dD_2 + (\bar{\alpha}_{i1} - \hat{\alpha}_{i1})dV_1 = 0$$

since $dS_2 = -dD_2$ from (9) and thus $dV_1 = 0$.

[12] When the capital market is financially complete, changes in the capital structure of a firm will not alter the investors implicit prices for $\theta$.

[13] The analysis here is analogous to that of Hayne Leland (1973), Theorem V.

All investors therefore are indifferent to a purely financial transaction. The second and third facets of the M-M theorem as stated in the introduction then hold with default risk for fixed-scale firms when the capital market is financially complete.[14]

If the debt is used to finance an increase in the scale of the firm instead of being distributed to shareholders, an initial shareholder ($\bar{\alpha}_{i2}>0$) can be shown to prefer an increase in debt-financed scale if and only if it increases the value of the equity of the firm.[15] This may be demonstrated by substituting the change in investor borrowing (lending)

$$(12') \quad d\hat{m}_i = (\bar{\alpha}_{i2} - \hat{\alpha}_{i2})dS_2 - \hat{\beta}_{i2}dD_2$$

into (11) to obtain

$$(14) \qquad \frac{dE_iU_i}{E_iU_i'} = \bar{\alpha}_{i2}dS_2$$

Since the preferences of all initial shareholders depend only on the sign of $dS_2$, all those shareholders will be unanimous with respect to their preferences for debt-financed investment and will prefer that debt-financed investment if and only if it increases the value of their equity holdings.

The change in the value of the equity of the firm resulting from debt-financed investment may be determined by rewriting (9) as

$$(15) \quad dS_2 = (k'(M_2)V_2/k(M_2) - 1)dD_2$$

The initial shareholder will prefer an increase (decrease) in debt if and only if the value of the firm is greater (less) than the total product divided by the marginal product $k'(M_2)$. The preferred debt level will be that at which $V_2=k(M_2)/k'(M_2)$, and the level of debt that yields this equality will maximize the value of the equity of the firm[16] as well as the expected utility of all initial shareholders. Acting so as to increase the value of the equity of the firm will thus be in the interests of shareholders. Since the optimality condition in (15) is independent of shareholder characteristics, the firm may act to maximize the expected utility of its shareholders without any knowledge of its shareholders except that they are expected utility maximizers.

### C. *Investor Preferences in an Incomplete Market*

If the capital market is neither scale nor financially complete, a purely financial transaction or debt-financed investment will create a new security and a new capital market equilibrium will result. The implicit prices are then not independent of the firm's financing and the above results do not obtain. Further results can be obtained however by substituting (4) and (5) into (10) to obtain[17]

$$(16) \quad \begin{aligned} \frac{dE_iU_i}{E_iU_i'} = {} & \hat{\beta}_{i2}(D_2k'(M_2)dM_2/k(M_2) - dD_2) \\ & + \hat{\alpha}_{i2}(S_2k'(M_2)dM_2/k(M_2) - dS_2) \\ & + \bar{\alpha}_{i2}dS_2 + \bar{\alpha}_{i2}dD_2 \\ & + (\hat{\beta}_{i2} - \hat{\alpha}_{i2})[R_2dD_2 + D_2dR_2 \\ & - R_2D_2k'(M_2)dM_2/k(M_2)] \\ & \cdot \int_{\theta^*}^{\infty} \omega_i(\theta)d\theta \end{aligned}$$

For fixed-scale firms and a purely financial transaction in a financially incomplete capital market, $dM_2=0$, and the expression in (16) is,[18] using $dS_2=dV_2-dD_2$,

$$\begin{aligned} \frac{dE_iU_i}{E_iU_i'} = {} & (\bar{\alpha}_{i2} - \hat{\alpha}_{i2})dV_2 \\ & + (\hat{\beta}_{i2} - \hat{\alpha}_{i2})((R_2-1)dD_2 + D_2dR_2) \\ & \cdot \int_{\theta^*}^{\infty} \omega_i(\theta)d\theta \end{aligned}$$

The preferences of a shareholder will not in

[14] Milne (1975) has provided a similar analysis of the M-M theorem.

[15] Since the scale of firm 1 does not change, $dV_1=0$.

[16] The preferred level of debt-financed investment will be less than that which maximizes the total value of the firm.

[17] For this analysis a capital market consisting only of the securities of one firm and the risk-free asset is considered. If more securities are available, the investor's preferences also depend on the changes in the return and in the values of those securities.

[18] The example given by Hagen involves such an incomplete capital market for which $dV_2/dD_2=dV_1/dD_2 \neq 0$.

general be in accord with the shareholder's gain or loss $((\bar{\alpha}_{i2}-\hat{\alpha}_{i2})dV_2)$ resulting from the change in the value of the firm. If, however, $\hat{\beta}_{i2}=\hat{\alpha}_{i2}$, preferences will be in accord with changes in value only if the investor will sell $(\bar{\alpha}_{i2}-\hat{\alpha}_{i2}>0)$ some of his equity holdings.[19]

For debt-financed investment with no default risk or for the case in which investors hold equal final proportions $(\hat{\alpha}_{i2}=\hat{\beta}_{i2})$ of the bonds and equity of the firm, the expression in (16) is[20]

$$\text{(17)} \quad \frac{dE_iU_i}{E_iU_i'} = \bar{\alpha}_{i2}dS_2 + \hat{\alpha}_{i2}(V_2k'(M_2)dM_2 / k(M_2) - dV_2)$$

The preferences of an investor for debt-financed investment then depend on both the change in wealth due to initial holdings and the marginal gain on the new equity holdings as represented by the last term in (17). This need not, however, be coincident with an increase in the value of the equity of the firm.[21] The model used by Smith involves an incomplete financial market and hence unanimity results do not obtain, in general, as indicated in the following section.

## III. A Paid-In Capital Model

Smith was perhaps the first to consider investor preferences for debt-financed investment. In his model the scale $M_2$ of the firm is obtained by the funds $V_2=(S_2+D_2)$ supplied by investors from their initial wealth (i.e., $\bar{\alpha}_{i2}=0$). Their preferences for debt-financed investment when there is no default risk is then determined from (17) as

$$\text{(18)} \quad \frac{dE_iU_i}{E_iU_i'} = \hat{\alpha}_{i2}(\epsilon - 1)dM_2$$

where $\epsilon=d\ log\ (k(M_2))/d\ log\ (M_2)$ is the elasticity of the total product with respect to paid-in capital. An investor prefers an increase (decrease) in the scale of the firm if the elasticity is greater (less) than one. With constant returns to scale the investor is indifferent to changes in capital structure as indicated by Smith in his Theorem 1. In the terminology utilized in this paper, the capital market in Smith's model is scale and financially complete if $\epsilon=1$ for all $M_2$. If $\epsilon\neq 1$, changes in scale introduce a new security (pattern of return) into Smith's capital market even if there is no default risk, so the capital market is incomplete.

With default risk the capital market is incomplete in Smith's model, and an increase in debt is preferred if and only if[22]

$$\text{(19)} \quad \frac{dE_iU_i}{E_iU_i'} = (\hat{\beta}_{i2} - \hat{\alpha}_{i2})A > 0$$

where

$$A = (D_2dM_2/M_2 - dD_2) + (R_2dD_2 + D_2dR_2 - R_2D_2dM_2/M_2)\int_{\theta^*}^{\infty} \omega_i(\theta)d\theta$$

Consequently, if there is default risk, the investor under Smith's assumptions will be indifferent to changes in the debt level of the firm only if either $\hat{\beta}_{i2}=\hat{\alpha}_{i2}$ or $A$ equals zero. Letting the investor's debt holdings be $D_{i2}^0=\hat{\beta}_{i2}D_2$ and equity holdings be $S_{i2}^0=\hat{\alpha}_{i2}S_2$, the investor with $\hat{\alpha}_{i2}=\hat{\beta}_{i2}$ is indifferent to a change in the debt level if $D_{i2}^0/S_{i2}^0=\mu$, where $\mu$ is the debt-equity ratio of the firm. Thus if the firm's debt-equity ratio is equal to the ratio of the investor's holdings of the firm's debt and equity, the investor is indifferent to changes in the debt level of the firm. This is in accord with Smith's Theorem 2 for the case of stochastic constant returns to scale. If, however, $\mu\neq D_{i2}^0/S_{i2}^0$, the preference of the individual cannot be predicted without

[19] If all investors have identical expectations and all have quadratic utility functions, for example, every investor will hold the same proportion of all securities. Under these strong assumptions the implicit prices are independent of the change in the capital structure and $dV_2=0$.

[20] With a scale and financially complete capital market the last term in (17) is zero from (12).

[21] Additional results can be obtained by analyzing the capital market at an *ex post* equilibrium in which investors already hold optimal portfolios; i.e., $\bar{\alpha}_{i2}=\hat{\alpha}_{i2}$ and $\bar{\beta}_{i2}=\hat{\beta}_{i2}$ where the bonds are considered as having been retraded. Leland (1974) and Ekern (1974) have analyzed such a case and have developed unanimity results.

[22] The condition in (19) is obtained by letting $dM_2=dD_2+dS_2$, $S_2=M_2-D_2$, and simplifying.

knowledge of his portfolio holdings, utility function, and probability assessments. A shareholder will prefer an increase (decrease) in the firm's debt if

$$(20) \qquad \mu > (<) D_{i2}^{0}/S_{i2}^{0} \quad \text{and} \quad A < 0$$

$$\text{or} \qquad \mu < (>) D_{i2}^{0}/S_{i2}^{0} \quad \text{and} \quad A > 0$$

The result given by Smith in Theorem 2 differs from that in (20) because he stated the preferences in terms of the debt-equity ratio instead of the debt level and because he assumed that the nominal interest rate on the firm's bonds was independent of the debt-equity ratio.

## IV. Conclusions

The appropriate criterion for financial and investment decisions is to take those decisions that are desired by shareholders. Determining what actually is in the interests of shareholders is in general difficult, since an investor's preferences depend on his utility function, probability assessments, and portfolio holdings. When, however, the proposed financing or investment by a firm does not alter the available returns in the capital market, the implicit prices for every investor will be unaffected by the proposals, and all shareholders will have unanimous preferences regarding such proposals. All investors then are indifferent to an alteration in the capital structure of a fixed-scale firm, and the value of that firm is independent of its capital structure. For variable-scale firms all shareholders prefer an increase in debt-financed investment if and only if it increases the value of the equity of the firm. In this case the firm may act in the best interests of its shareholders by maximizing the value of the equity of the firm. The second and third facets of the M-M theorem thus obtain.

If the financing–investment alters the space of available returns in the capital market, the implicit prices will depend on the firm's decisions except under certain special assumptions such as those that lead to separation. When the implicit prices are not independent of the firm's proposals, a purely financial transaction by one fixed-scale firm will change the value of all fixed-scale firms in a risk class by the same amount, since those firms must have the same value. With a variable scale the value of a firm will depend both on its financing and its investment opportunities. In such cases where the implicit prices are not independent of the decisions of the firm, the firm is without a guide as to how to work in the best interests of shareholders, since the best interests of shareholders are not in general coincident with changes in the value of the equity of the firm. Also, shareholders will not in general be in agreement with respect to their preferences for the proposals of a firm, and the firm is then without an unambiguous criterion for its financial and investment policies.

## REFERENCES

D. P. Baron, "Default Risk, Homemade Leverage, and the Modigliani-Miller Theorem," *Amer. Econ. Rev.*, Mar. 1974, *64*, 176–82.

D. Cass and J. E. Stiglitz, "The Structure of Investor Preferences and Asset Returns, and Separability in Portfolio Allocation: A Contribution to the Pure Theory of Mutual Funds," *J. Econ. Theory*, June 1970, *2*, 122–60.

P. A. Diamond, "The Role of a Stock Market in a General Equilibrium Model with Technological Uncertainty," *Amer. Econ. Rev.*, Sept. 1967, *57*, 759–76.

S. Ekern, "Some Aspects of Firms' Decision Making in an Economy with Incomplete Capital Market," *Swedish J. Econ.*, Mar. 1974, *76*, 117–30.

———, "On the Theory of the Firm in an Economy with Incomplete Markets: An Addendum," *Bell J. Econ.*, Spring 1975, *6*, 388–93.

——— and R. Wilson, "On the Theory of the Firm in an Economy with Incomplete Markets," *Bell J. Econ.*, Spring 1974, *5*, 171–80.

K. P. Hagen, "Default Risk, Homemade Leverage, and the Modigliani-Miller Theorem: Note," *Amer. Econ. Rev.*, Mar. 1976, *66*, 199–203.

H. E. **Leland**, "Production Theory and the Stock Market," *Bell J. Econ.*, Spring 1974, *5*, 125–44.

———, "Capital Asset Markets, Production and Optimality: A Synthesis," techn. rep. no. 715, Inst. Mathemat. Stud. Soc. Sciences, Stanford Univ., Dec. 1973.

F. **Milne**, "Choice over Asset Economics: Default Risk and Corporate Leverage," *J. Financial Econ.*, June 1975, *2*, 165–85.

———, "Corporate Investment and Finance Theory in Competitive Equilibrium," *Econ. Rec.*, Dec. 1974, *50*, 511–33.

F. **Modigliani** and M. H. **Miller**, "The Cost of Capital, Corporation Finance, and the Theory of Investment," *Amer. Econ. Rev.*, June 1958, *48*, 261-97.

R. **Radner**, "A Note on Unanimity of Stockholders' Preferences Among Alternative Production Plans: A Reformulation of the Ekern-Wilson Model," *Bell J. Econ.*, Spring 1974, *5*, 181–84.

V. L. **Smith**, "Default Risk, Scale, and the Homemade Leverage Theorem," *Amer. Econ. Rev.*, Mar. 1972, *62*, 66–76.

J. E. **Stiglitz**, "A Re-Examination of the Modigliani-Miller Theorem," *Amer. Econ. Rev.*, Dec. 1969, *59*, 784–93.

# The Xistence of X-Efficiency

*By* GEORGE J. STIGLER*

Harvey Leibenstein called attention in an influential article (1966) to a source of economic inefficiency which was given the awful name of X-[in]efficiency. He cited studies in which misallocations of resources due to monopoly or tariffs had trifling social costs, whereas simple failure to attain the production frontier apparently led to social losses of a vastly greater magnitude. I propose to argue that this type of inefficiency can usefully be assimilated into the traditional theory of allocative inefficiency.

It is a question (to be discussed below) whether one ascribes failures to reach the ultimate limits of output from given inputs in any state of technology to inadequacy of knowledge alone, or adds also inadequate "motivation." Leibenstein (1966) separates the two:

> It is obvious that not every change in technique implies a change in knowledge. The knowledge may have been there already, and a change in circumstances induced the change in technique. In addition, knowledge may not be used to capacity just as capital or labor may be underutilized. More important, a good deal of our knowledge is vague. [pp. 404–05]

He ascribes increases in X-efficiency to 1) increases in motivational efficiency—workers are stimulated by incentive pay, or management by competition or other adversities; and 2) improvements in the inefficient markets for knowledge. I shall first deny the propriety of treating changes in motivation as a source of changes in output, and then proceed to the question of knowledge.

### I. Motivational Losses

Leibenstein (1973) has emphasized that X-inefficiency arises largely from losses of output due to motivational deficiencies of resource owners:

> [F]or the same set of human inputs purchased and the same knowledge of production techniques available to the firm, a variety of output results are possible. If individuals can choose, to some degree, the APQT bundles [choice of Activity, Pace, Quality of work, Time spent] they like, they are unlikely to choose a set of bundles that will maximize the value of output. [p. 768]

If management seeks to impose output-maximizing *APQT* bundles on the workers, indeed, these assignments of tasks would likely be " . . . less efficient than those that individuals would choose themselves under an acceptable set of [managerial] restraints" (p. 769).

In this case, and in every motivational case, the question is: what is output? Surely no person ever seeks to maximize the output of any one thing: even the single proprietor, unassisted by hired labor, does not seek to maximize the output of corn: he seeks to maximize utility, and surely other products including leisure and health as well as corn enter into his utility function. When more of one goal is achieved at the cost of less of another goal, the increase in output due to (say) increased effort is not an increase in "efficiency"; it is a *change* in output.

The concept of motivational efficiency seems to extend also to the task of getting a "predetermined output" from hired factors (see Leibenstein, 1966, pp. 408, 412). There are important and pervasive problems in all contracts between people, in seeking the fulfillment of the reciprocal contractual promises, and substantial resources are necessary to enforce the agreements (see Armen Alchian and Harold Demsetz). Both the avoidance of unpleasant tasks and the enforcement activity designed to curtail this avoidance can be carried on to the utility-maximizing degree and generate no inefficiency in producing utility. Output and utility would be larger if resources were not

* Graduate School of Business, University of Chicago.

necessary to the enforcement of contracts, but output and utility would also be larger if water boiled at 180°F or a day had 25 hours. New techniques of contract enforcement may be as productive as other improvements of technology.

Thus X-inefficiency attributed to motivational factors characterizes as inefficiency either the existence and pursuit of other desired outputs or the expenditure of resources required for the optimal enforcement of contracts. This tunnel vision of output seems entirely unrewarding: it imposes one person's goal upon other persons who have never accepted that goal. There is no waste in this sort of X-inefficiency: waste is a foregone product that could be acquired for less than its cost.

Leibenstein achieves much of the importance of motivation in X-efficiency by that ancient and powerful scientific technique, definition. When he copiously illustrates that "changes in incentives will change productivity per man" (1966, p. 401), he is assigning motivation an independent role whereas ordinary economic language would classify the methods of remuneration of employees as a part of the state of technology. Again, when an Egyptian petroleum refinery becomes more "efficient" with a change of management, we are told that "It is quite possible that had the motivation existed in sufficient strength, this change could have taken place earlier" (1966, p. 398). Potential motivation could indeed rewrite all history: if only the Romans had tried hard enough, surely they could have discovered America. (Thus motivation can be invoked to explain every unperformed task that is physically possible, no matter how unrewarding.) We may sympathize with Leibenstein's desire to associate his X-efficiency with economic behavior, but this shotgun marriage is not fertile.

## II. The State of Technology

The near-universal tradition in modern economic theory is to postulate a maximum possible output from given quantities of productive inputs—this is *the* production function—and to assert that each firm operates on this production frontier as a simple corollary of profit or utility maximization. The merit of this conventional tradition is also its demerit: it eliminates the problem of the choice of technology.

Alfred Marshall followed an entirely different approach, and it is remarkable that he had virtually no followers. He proposed to characterize production possibilities by the *average* outputs obtained from given inputs, and in particular labelled the user of this average relationship the Representative Firm:

> We shall have to analyse carefully the normal cost of producing a commodity, relatively to a given aggregate volume of production; and for this purpose we shall have to study *the expenses of a representative producer* for that aggregate volume. On the one hand we shall not want to select some new producer just struggling into business, who works under many disadvantages, and has to be content for a time with little or no profits, but who is satisfied with the fact that he is establishing a connection and taking the first steps towards building up a successful business; nor on the other hand shall we want to take a firm which by exceptionally long-sustained ability and good fortune has got together a vast business, and huge well-ordered workshops that give it a superiority over almost all its rivals. But our representative firm must be one which has had a fairly long life, and fair success, which is managed with normal ability, and which has normal access to the economies, external and internal, which belong to that aggregate volume of production; account being taken of the class of goods produced, the conditions of marketing them and the economic environment generally. [p. 317, and Bk. IV, ch. 13]

Marshall suggested two causes of variation among firms in costs of a given output: the age of the firm (which he emphasized), and variations in entrepreneurial capacity. Strictly speaking, the latter element (the departure from "normal ability") is inappropriate: differences in quality of an input do not lead to differences in outputs from given inputs.

The reason Marshall's approach was not adopted by the science is lucidly presented in the leading attack that was made on the

representative firm by Lionel Robbins. In a once-famous essay, Robbins argued persuasively that when costs of firms differed because of quality of entrepreneurs (or other inputs), the differences in productivity would be reflected in differences in profits (or other input prices). Just as differences in efficiency of workers are reflected in their wages, so differences in entrepreneurial skills (including the choice of technology) will be reflected in their "profits." He states:

> There is no more need for us to assume a representative firm or representative producer, than there is for us to assume a representative piece of land, a representative machine, or a representative worker. [p. 393]

Robbins was of course correct: the Representative Firm is not needed to reconcile the existence of differences among entrepreneurs with the existence of stable competitive equilibrium.

What one may lament, however, is the failure of Robbins and Leibenstein, and all of us in between, to recognize the problem of determining which technologies will be used by each firm (and, for that matter, each person). The choice is fundamentally a matter of investment in knowledge: the costs and returns of acquiring various kinds and amounts of technological information vary systematically with various characteristics of a firm: its size, the age of its present capital assets, the experience of its managers, the prospects of the trade. No attention has been paid by economists to the analysis of the optimal amount of technological knowledge that a firm should possess. Leibenstein deserves credit for reviving this Marshallian question, but his attention to X-inefficiency as the explanation is an act of concealment: it simply postulates the differences in technology among firms which should be explained.

### III. The Interpretation of Output Differences

We observe two farmers with reasonably homogeneous land and equipment, who nevertheless obtain substantially different amounts of corn. We measure this corn output over some period of time to reduce the effects of stochastic variation (i.e., unenumerated inputs such as weather). The observed variation is due, perhaps, to differences in knowledge, including the knowledge of technology or the knowledge of how far to carry the application of each productive factor. The farmers will differ in the cost of learning new things or the expected returns from new knowledge—one may be planning to leave agriculture shortly—so they "rationally" devote different amounts of resources to acquiring knowledge. Or one is simply more intelligent than the other, and learns more quickly or thinks more precisely (for example, makes fewer mistakes in arithmetic).

The effects of these variations in output are all attributed to specific inputs, and in the present case chiefly to the differences in entrepreneurial capacity. In neoclassical economics, the producer is always at a production frontier, but his frontier may be above or below that of other producers. The procedure allocates the foregone product to some factor, so in turn the owner of that factor will be incited to allocate it correctly.

Leibenstein does not attempt to understand the allocation of "inefficient" resources, and hence does not see the necessity for attributing his X-inefficiency to specific inputs. Just as automobile accidents are palpable inefficiencies to many people so X-inefficiency is a palpable inefficiency to Leibenstein. But accidents and "inefficiencies" are associated with returns as well as costs, and a useful theory must take both sides of their roles into account.

Indeed, Leibenstein's apparatus does not allow him to analyze effectively concrete economic problems. Consider his argument that monopoly is less efficient than competition. To reach this result, he must assume that 1) monopolists do not maximize profits, and 2) competitors are driven closer to "the" minimum costs by the entry of new rivals, some of whom are efficient, by a Darwinian process. The first assumption is an abandonment of formal theory, and one which we shall naturally refuse to accept until we are given a better theory. It "solves" the question of the effect of monopoly on efficiency without argument or evidence. The latter assumption of competitive selection coolly

ignores the problem of general equilibrium (where do the driven-out entrepreneurs go?, and where do the efficient entrepreneurs come from?), and fails to demonstrate (or even to argue) that inflows and outflows of entrepreneurs of various qualities will converge on a high-efficiency equilibrium in each competitive industry.

Earlier I defined waste as the situation in which foregone products could be obtained for less than they cost. Waste can arise *ex post* because *ex ante* plans rested upon erroneous predictions. This type of waste is unavoidable, although its magnitude is subject to control. Waste can also arise in the absence of uncertainty if the economic agent is not engaged in maximizing behavior. Unless one is prepared to take the mighty methodological leap into the unknown that a nonmaximizing theory requires, waste is not a useful economic concept. Waste is error within the framework of modern economic analysis, and it will not become a useful concept until we have a theory of error.

REFERENCES

**A. A. Alchian** and **H. Demsetz**, "Production, Information Costs, and Economic Organization," *Amer. Econ. Rev.*, Dec. 1972, *62*, 777–95.

**H. Leibenstein**, "Allocative Efficiency vs. 'X-Efficiency'," *Amer. Econ. Rev.*, June 1966, *56*, 392–415.

———, "Competition and X-Efficiency: Reply," *J. Polit. Econ.*, May/June 1973, *81*, 765–77.

**A. Marshall**, *Principles of Economics*, 8th ed., London 1948, 317.

**L. Robbins**, "The Representative Firm," *Econ. J.*, Sept. 1928, *38*, 387–404.

# A Note on Quantitative Restrictions and Capital Mobility

By RODNEY E. FALVEY*

The link between international factor mobility and commodity trade has received some attention in recent years. In the main, this analysis has served to extend and "tighten" Robert Mundell's original demonstration of the basic substitutability between goods and factor movements in the standard trade model,[1] and to expand the discussion to examine the interaction between commodity and factor taxation.[2] In the absence of impediments to factor movements (physical restrictions or differential taxation on foreign and domestic earnings), any import tariff is prohibitive, in the sense that the factor flows it induces will eliminate international commodity exchange.[3]

Unless both commodity and factor taxation are possible, it appears a country has little influence over the composition of its trade or domestic production. When factors are immobile, the optimal policy to achieve some desired production level other than the free trade level is to use the appropriate production tax-subsidy scheme. With factor mobility, however, this runs into the same problems as the tariff and would eventually force the country to specialize in the protected industry. But tariffs (or price restrictions) represent but one major form of commercial policy. Previous analysis has neglected the other—quantitative restrictions.

The object of this note then is two-fold. First to demonstrate that quantitative restrictions on commodity trade will permit a country to influence its trade and production patterns, despite factor mobility, and without requiring distortions in both factor and commodity markets. Second, while doing this, to demonstrate the basic differences in the adjustment processes under these two commercial policy regimes. Although, in a static framework they are essentially equivalent, once factor mobility is introduced the distinction between a price and a quantity restriction becomes important. This last result is quite general and readily extends beyond the confines of the structure examined here. I begin by briefly considering the effects of an import tariff on trade and production in this system, and then go on to compare this with the effects of an import quota.

Suppose the world is composed of two countries (home and foreign), each of which produces two goods (the home exportable $X_e$ and the home importable $X_m$), through the services of two factors of production (labor and capital). Assuming identical linear homogeneous technologies in each country, let the factor endowments be such that the initial free trade equilibrium implies factor price equalization. Thus, although capital is assumed internationally mobile, there exists no incentive in the form of a return differential for it to move. There are two aspects of this equilibrium which are important for future results: First, the allocation of capital between the two countries which would give this equilibrium is not unique.[4] Second, within a certain range,[5] shifting capital from

* Assistant professor of economics, Virginia Polytechnic Institute and State University. I am indebted to Rudiger Dornbusch, Michael Mussa, and a referee for comments and suggestions.

[1] See Mundell and, for example, Frank Flatters, Douglas Purvis, and Melvyn Krauss. Certain ambiguities can arise if both factors are internationally mobile, however. First, determining what it is that defines a country (usually its endowment), and second, if tastes are different in the two countries, determining who it is that holds these tastes.

[2] See Ernest Nadel. Particularly when the assumption of identical technologies is relaxed—see Ronald Jones and John Chipman.

[3] A one-way commodity flow will still occur, however, if foreign earnings are repatriated. Alternatively, the import restriction could force specialization in production of the importable.

[4] There is a range of "endowment allocations" for which the same total world production, relative prices and factor returns could occur. See Robert Warne.

[5] Depending on how close each is to specialization initially. Again see Warne.

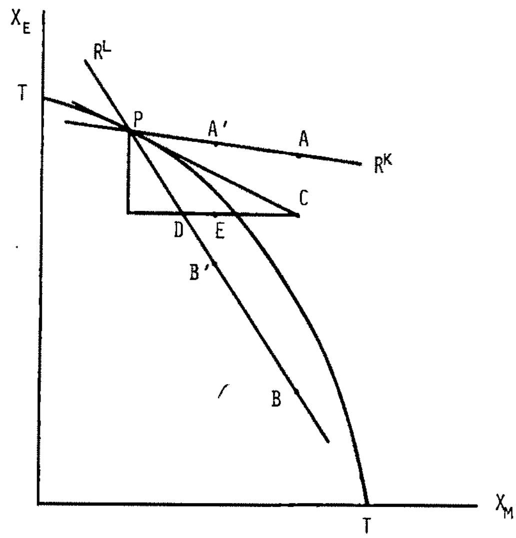



FIGURE 1

one country to the other simply changes the location of production without altering its total, because of linear homogeneity.

For convenience, it is assumed both countries own their entire initial domestic capital stocks, and, in what follows, that any foreign earnings on capital will be repatriated in full. This initial equilibrium is demonstrated for the home country in Figure 1. As shown, $TT$ is the domestic production-possibility frontier for its initial endowments of factors; $P$ represents the free trade production point; and $C$ the consumption point, when the terms of trade are represented by the tangent at $P$.

Now let the home country place a tariff on its imports with any tariff proceeds being redistributed in a neutral manner. Imagine, for a moment, capital is immobile. Then, barring a Metzler Paradox,[6] the domestic relative price of the importable would rise and the home terms of trade improve. Thus, relative prices would move in opposite directions in the two countries, and the real return to the factor used relatively more intensively in the production of the home importable would increase at home and decline abroad.[7] Conversely for the other factor.

Should capital's domestic return tend to rise, the home country will experience a capital inflow, otherwise a capital outflow. In either event, the force of the resulting capital movement is toward expanding domestic production of importables, independent of factor intensities. Foreign production of their exportable correspondingly contracts. This capital flow will continue as long as the international differential in rentals exists—that is, until either the home country becomes specialized in the production of the importable, or commodity prices return to their original levels.[8] Neglecting specialization, the price and rental differential persists until production changes have eliminated trade in the importable (whence the tariff becomes redundant). At this point world prices, incomes, and hence, desired consumption will have returned to their initial levels (i.e., $C$ in Figure 1).

Consider the two alternative intensity assumptions.

1) The importable is capital intensive. At constant commodity prices, the capital inflow following the tariff shifts the corresponding production point along the Rybczynski line $R^K$, until point $A$ is reached. At $A$, domestic production of the importable is just sufficient to satisfy desired consumption at free trade prices and thus imports cease. The quantity $CA$ of the exportable represents the income to foreign-owned capital.

2) The importable is labor intensive. At constant commodity prices, capital leaves shifting the production point down the Rybczynski line $R^L$, until $B$ is reached. Here the home country is self-sufficient in the importable and receiving $BC$ of the exportable as income from its foreign investments.[9]

[6] A possibility under a tariff regime but precluded by an import quota—see the author. The results are unaffected by this possibility however.

[7] The Stolper-Samuelson Theorem: see Wolfgang Stolper and Paul Samuelson.

[8] Since the thrust of the capital movement is toward reduced foreign production of their exportable, trade will be eliminated long before the foreign country could become specialized.

[9] If the tariff did not apply to earnings repatriated in units of the importable, the new production equilibrium could be at $D$, where the home country is self-sufficient in the exportable, or at any point on $DB$.

Basically the reason the solution goes this far is as follows. When both countries are nonspecialized in this model, factor returns depend on commodity prices and the technology alone, and thus are unaffected by the capital movements in themselves. The shift in capital from one country to the other merely serves to shift the location of production without altering its total. What this result illustrates is the basic substitutability between commodity and factor movements in this model. Any tariff, by creating a fixed price distortion between the two countries, prompts counteracting capital flows which eliminate trade.

A tariff alone is not a flexible instrument in these circumstances. With unrestricted capital movements, any tariff is prohibitive. There are but two production levels the country can choose—"free trade" or "no trade." A dilemma then arises if our country desires production at some intermediate point.[10] Can this be achieved without distorting both factor and commodity markets? The solution lies in quantitative restrictions. While a tariff creates a constant price differential between the two countries as long as any trade in the importable occurs, an import quota distorts prices only to the extent trade is restrained by the quota. By imposing the appropriate import quota, our country can shift production of the importable to any level between the free trade and no trade production levels it chooses. For example, consider a quota limiting imports to the level shown by distance *EC* in Figure 1. In the absence of capital mobility, this forces the domestic relative price of the importable to rise above its free trade level, expanding production and discouraging demand to the extent necessary to reduce imports to the quota level. Conversely, the relative price of the home importable will fall in the foreign country. Again, we have two alternative pressures on factor returns, depending on which factor is used intensively in production of the importable. In either case, domestic import competing production will expand as a consequence of the resulting capital movement. Here the difference between a price and a quantity distortion becomes important. As long as the quota is binding, it separates the markets for the home importable in two countries and the location of production is no longer immaterial.

As production of the home importable shifts with the capital flow from the foreign to the home country, we get an excess supply at home and an excess demand abroad at the initial postquota prices. The relative price of the domestic importable tends to fall at home and rise abroad, pulling the real returns to capital closer together in the two countries. Thus, unlike the tariff, under a quota regime the flow of capital is able to influence the differential in the real returns to capital through the resulting shifts in the location of production.

As before, this capital flow continues until returns are equalized. Again, barring specialization, this occurs when relative prices return to their former levels. However, unlike the tariff, imports are not eliminated but simply reduced to the level established by the quota where the price and return differentials, and hence the incentive for further capital movements are eliminated. In Figure 1, we are now back at free trade prices, produced either at $A'$ or $B'$.

This demonstrates that quantitative restrictions retain their "flexibility" in forcing production adjustments, despite the possibility of capital movements.[11] An import tariff and quota pair which will be equivalent with capital immobile are no longer equivalent when capital can move. The rigid price differential forced by the tariff persists as long as trade in the importable exists. With the quota, however, a price differential exists only to the extent trade is limited by the quota. So any tariff has as its equivalent the prohibitive quota.

[10] Perhaps in pursuit of some "noneconomic objective." For a discussion of such objectives, see Jagdish Bhagwati. Note I am assuming our country is unconcerned about capital movements per se.

[11] Note, however, that while the quota is a flexible instrument with respect to the national origin of production, it remains ineffective with respect to factor rentals and goods prices. The main thrust of Mundell's argument, that capital mobility frustrates attempts to alter prices and rentals, is unchanged.

The adjustment process also differs under the two regimes, since the link between product, and hence factor, prices in the two countries is fixed under the tariff but free to vary with the capital movements under the quota. Thus, as the capital flows occur, they exert a direct effect reducing the return differential under the quota but not under the tariff. This property will carry over to more complex structures where capital's return depends on both relative prices and the capital "endowment." Now capital flows will have a direct influence on the return to capital, independent of any change in relative prices that may occur. In addition the pattern of relative prices in the two countries converging with the capital flow under the quota, yet remaining separated by a fixed wedge under the tariff, will persist. Given this, in such a more general structure one expects smaller flows and production shifts under the quota regime.

## REFERENCES


**J. Bhagwati,** "The Theory and Practice of Commercial Policy: Departures from Unified Exchange Rates," *Princeton Special Papers Int. Econ.*, no. 8, Jan. 1968.

**J. Chipman,** "International Trade with Capital Mobility: A Substitution Theorem," in J. N. Bhagwati et al., eds., *Trade, Balance of Payments and Growth: Papers in International Economics in Honor of Charles P. Kindleberger*, Amsterdam 1971, 201–37.

**R. E. Falvey,** "A Note on the Distinction Between Tariffs and Quotas," *Economica*, Aug. 1975, *42*, 319–26.

**F. Flatters,** "Commodity Price Equalization: A Note on Factor Mobility and Trade," *Amer. Econ. Rev.*, June 1972, *62*, 473–76.

**R. W. Jones,** "International Capital Movements and the Theory of Tariffs and Trade," *Quart. J. Econ.*, Feb. 1967, *81*, 1–38.

**M. B. Krauss,** "Commodity Trade and Factor Mobility," *Amer. Econ. Rev.*, Sept. 1974, *64*, 797–801.

**R. A. Mundell,** "International Trade and Factor Mobility," *Amer. Econ. Rev.*, June 1957, *49*, 321–35.

**E. Nadel,** "International Trade and Capital Mobility," *Amer. Econ. Rev.*, June 1971, *61*, 368–79.

**D. D. Purvis,** "Technology, Trade and Factor Mobility," *Econ. J.*, Sept. 1972, *82*, 991–99.

**W. Stolper and P. Samuelson,** "Protection and Real Wages," *Rev. Econ. Stud.*, Nov. 1961, *9*, 58–73.

**R. D. Warne,** "National Endowments and Efficient World Production, A Geometrical Treatment," *Manchester Sch. Econ. Soc. Stud.*, Sept. 1972, *40*, 313–17.

# Black-White Differences in Returns to Schooling: Some New Evidence

By Charles Link, Edward Ratledge, and Kenneth Lewis*

During the 1960's it was generally believed that policies of providing additional education would go a long way toward equalizing the large income differentials between blacks and whites. But in light of recent studies on the subject, serious doubt has been cast on the sanguine view toward education's ability to accomplish income equality (see Giora Hanoch, Randall Weiss, and Lester Thurow). While most research has centered on returns to *years of education*, Christopher Jencks has gone even further to argue that the *quality* of education has a negligible part to play in the process. Recently, however, the studies by Leonard Weiss and Jeffrey Williamson, Link and Ratledge, and Finis Welch have yielded more optimistic findings regarding the rôle of educational quality on black-white earnings differences.[1]

The present note deals specifically with the recent article by Welch. He presents evidence suggesting that blacks for the past several years have finally begun to receive monetary benefits from education commensurate with those of whites. Most importantly, he argues that this change is due to an increase in the quality of black education. Welch observes in his earnings equations a decline in the coefficient of the years of schooling variable ($\hat{a}$) for older cohorts of blacks and whites in his samples.[2] More important from Welch's point of view, the rate of decline in ($\hat{a}$) is larger for blacks than whites, implying blacks are catching up. It is the catching up which Welch attributes to increased quality of education for blacks.

There are several potential interpretations of Welch's results. One is that life cycle or credential effects may be present that lessen the impact of schooling on earnings as one is out of school longer.[3] However, even if life cycle effects are nonexistent, two additional explanations of the greater dropoff for blacks are possible. On the one hand, changes in the quality of education may cause higher returns to an additional year of schooling for blacks than whites. In other words, Welch's "so-called" vintage effect is greater for blacks than whites Alternatively, the quality of education may be rising more quickly for blacks than for whites.[4]

The regression results presented below indicate that the vintage effect is only slightly larger for blacks, corroborating Welch's hypothesis that improved quality of black education is indeed responsible for blacks' relative income gains. Similar to Welch's findings, life cycle effects are nonexistent for the sample of young males analyzed below.

To gain insight into the issues posed

* Associate professor of economics; associate director, Division of Urban Affairs; and associate professor of economics; respectively, at the University of Delaware. We would like to thank the anonymous referee and George Borts for comments on earlier drafts of the paper. The research was made possible by a UNIDEL grant from the University of Delaware.

[1] James Morgan and Ismail Siragelden as well as Frank Stafford and George Johnson found a positive association between average state expenditures per student in elementary and secondary education and subsequent earnings for whites.

[2] Welch utilizes samples of urban males from the 1 in 1000 sample of the 1960 Census and the 1966 Survey of Economic Opportunity.

[3] Thurow (1972) has argued that the impact of education on earnings does not arise from the cognitive skills imparted to an individual but is instead the result of the information it provides the employer about the trainability of the prospective employee. Thus as one gets further away from school, experience would become more important. Welch argues that life cycle effects are small and to support his claim makes cohort comparisons involving the direct effects in 1959 and 1966; within cohort groups he finds that the effects of schooling do not diminish over time.

[4] The three propositions discussed in the text are developed in a technical appendix available on request from the authors. Of course, the increases in the education coefficient for younger cohorts may reflect less discrimination in the labor market.

above, an earnings regression is estimated which directly incorporates a measure of elementary and secondary educational quality. The data are derived from the National Longitudinal Survey of the Labor Force which is being conducted by the Center for Human Resources Research at the Ohio State University under contract to the U.S. Department of Labor. The sample utilized consists of 1,562 males who were between the ages of 17 and 27 in 1969; and it includes 326 young blacks and 1,236 young white males who were out of school at least one year in 1969

The advantages of the data are threefold. First, they supply a proxy for the quality of elementary and secondary education received by each respondent. Second, blacks were oversampled by a factor of three, allowing greater precision in a statistical analysis of their labor force behavior. Finally, because the sample is composed of young males, we can examine the cohort of blacks which Welch argues has been the main beneficiary of increased educational quality.

The dependent variable is 1969 annual wages and salaries ($Y$). Explanatory variables included in the regression are: the natural logarithm of the number of weeks worked ($ln\ W$); potential labor market experience designated as age—education—5, subject to the constraint the useful experience starts at age 16 (*Exper*); experience squared $(Exper)^2$; an interaction between years of schooling ($S$) and the natural logarithm of expenditures per student as of 1968 in the district where the respondent went to high school (*ln Expend*); and an interaction between years of schooling ($S$) and experience (*Exper*). The two interactions allow us to determine whether or not Welch's "vintage" effect (i.e., $\partial\ ln\ Y/\partial S = f$ (quality)) and life cycle effects $(\partial\ ln\ Y/\partial S) = g$ (experience)) are present. A positive sign is predicted on the coefficient of $S$ x *ln Expend* and a zero or negative sign on the coefficient of $S$ x *Exper*.

The interaction between $S$ and *ln Expend* is employed instead of the additive variable years of schooling because units of education should be conceptually viewed as being the product of the years of schooling ($S$) and quality of education (*ln Expend*).[5] The unweighted schooling variable is in fact inappropriate unless the quality of a year of schooling is approximately the same for everyone in the sample.

TABLE 1—EARNINGS MODELS INCLUDING *ln* EXPENDITURES/STUDENT[a]

| Variable | White | Black |
|---|---|---|
| Weeks | .961[b] | .871[b] |
| | (.040) | (.059) |
| Experience | .290[b] | .149 |
| | (.052) | (.104) |
| Experience Square | −.015[b] | −.008[c] |
| | (.002) | (.004) |
| Schooling x Expenditures | .017[b] | .019[b] |
| | (.002) | (.005) |
| Schooling x Experience | −.003 | .000 |
| | (.003) | (.007) |
| Constant | 2.729 | 3.083 |
| Adjusted $R^2$ | .56 | .51 |

[a] Standard errors in parentheses.
[b] Significant at 1 percent level.
[c] Significant at 5 percent level.

Expenditures on education rose rapidly during the 1960's implying that our expenditure measure most likely overstates expenditures for young males educated prior to 1968. The coefficient of the schooling-expenditures interaction is likely to be biased downward. In other words, when these overstated expenditures (for those educated before 1968) are associated with *low* earnings in 1969, the observed relationship will tend to "wash out." If, as Welch (p. 901) suggests, expenditures per student on blacks have risen relative to whites, the bias will even be larger in the black equations.

The results of the earnings regressions are shown in Table 1.[6] The main coefficients of interest are those attached to the interactive variables $S$ x *ln Expend* and $S$ x *Exper*. Of

[5] Welch (p. 896) also recognizes this point, but data limitations prevented him from using the preferred formulation.

[6] Covariance analysis reveals that the structure of the black and white equations are significantly different at both the 1 and 5 percent levels of significance, even when we allow for differences in the black-white intercepts.

these only the former is significant at the 5 percent level. Most importantly the results in Table 1 confirm the existence of a vintage effect (i.e., $\partial \ln Y/\partial S = f$ (quality)). Second, the effect appears to be at least as large for blacks (.019) as whites (.017).[7] The data do not reflect life cycle effects, which is not surprising since the sample has a range of only eleven years of experience.

The major findings of the study indicate, as Welch suggested, that the great increase in returns to black versus white education is due to changes in the quality of education as opposed to differential vintage effects.[8] In addition, the interactive relationships for blacks exhibited between years and expenditures may partially explain the pessimistic finding of education's lack of impact on earnings of blacks found by researchers using the 1960 data. At the least the findings suggest that there is still a role for the quality of education to play in reducing black-white earnings differences. Of course a definitive statement on the issue awaits further research. However, even though Welch raised more questions than he answered, it appears that his major conclusions still hold and provide the basic hypothesis upon which future research can be based.

[7] The black-white differential is understated to the extent the black expenditure component of the interaction term reflects a larger downward bias compared to that of whites.

[8] The conclusions regarding quality remain when the sample is divided into groups of urban blacks and whites.

REFERENCES

**G. Hanoch**, "An Economic Analysis of Earnings and Schooling," *J. Hum. Resources*, Summer 1967, *2*, 310–19.

**C. Jencks et al.**, *Inequality: A Reassessment of the Effect of Family and Schooling for America*, New York 1972.

**G. Johnson and F. P. Stafford**, "Social Returns to Quantity and Quality of Education," *J. Hum. Resources*, Spring 1973, *8*, 139–55.

**C. Link and E. C. Ratledge**, "Social Returns to Quantity and Quality of Education: A Further Statement," *J. Hum. Resources*, Winter 1975, *10*, 78–89.

**J. Morgan and I. Sirageldin**, "A Note on the Returns to Quality of Schooling," *J. Polit. Econ.*, Sept./Oct. 1968, *76*, 1069–77.

**L. Thurow**, *Poverty and Discrimination*, Washington 1969.

———, "Education and Economic Equality," *Publ. Interest*, Summer 1972, *28*, 66–81.

**L. Weiss and J. G. Williamson**, "Black Education, Earnings, and Interregional Migration: Some New Evidence," *Amer. Econ. Rev.*, June 1972, *62*, 372–83.

**R. Weiss**, "The Effect of Education on the Earnings of Blacks and Whites," *Rev. Econ. Statist.*, May 1970, *52*, 150–59.

**F. Welch**, "Black-White Differences in Returns to Schooling," *Amer. Econ. Rev.*, Dec. 1973, *67*, 893–907.

**Center for Human Resources Research**, "National Longitudinal Survey of the Labor Force, Ohio State Univ.

# The Effect of Compensated Price Changes on Expenditures

*By* C. Robert Wichers*

After a compensated or overcompensated increase of the price of a commodity, the consumer buys fewer units of that commodity. Does he pay more for those fewer, more expensive units, or less? The Slutsky equation holds no answer (see John Hicks, p. 36). It will be shown, however, that an answer can be found through a slightly different definition of (over)compensation. Specifically, if a change of the price of a commodity is accompanied by (over)compensation *in terms of that commodity*, the consumer's expenditure for that commodity will change in a predictable direction (opposite to the direction of the price change, in fact). This is an empirically verifiable result.

Compensation or overcompensation in terms of commodities is of course less efficient in the welfare-theoretical sense than compensation or overcompensation in terms of money. This consideration plays no role here, however. The sole purpose of this paper is to establish an empirically verifiable result.

The result to be proved can be geometrically illustrated—quite easily so when money is treated as one of the utility-carrying commodities, less easily when money is not so regarded. (Either way, the result holds.) Only the easy case is sketched here. Figure 1 depicts a situation with two utility-bearing commodities, one of which is money. The individual's budget line is marked *I*, his initial bundle is $x^0$, and the equilibrium bundle is $x$. When $p_1$ increases, and a compensating quantity *of good 1* is added to the initial bundle $x^0$, the initial bundle becomes $\bar{x}^0$. At the same time, the budget line becomes the line marked *II*, and the equilibrium bundle becomes $\bar{x}$. As can be seen in Figure 1, the consumer now buys less of good 1 than before ($b\bar{x}<ax$) *and* pays less money than before ($b\bar{x}^0<ax^0$). (Compare the more conventional situation of Figure 2, where compensation is made in terms of money. Here, too, the compensated price increase induces the consumer to buy less of good 1 than before ($b\bar{x}<ax$), but it is not clear whether the consumer pays less money than before, or more, or exactly as much ($b\bar{x}^0 \lesseqgtr ax^0$).)

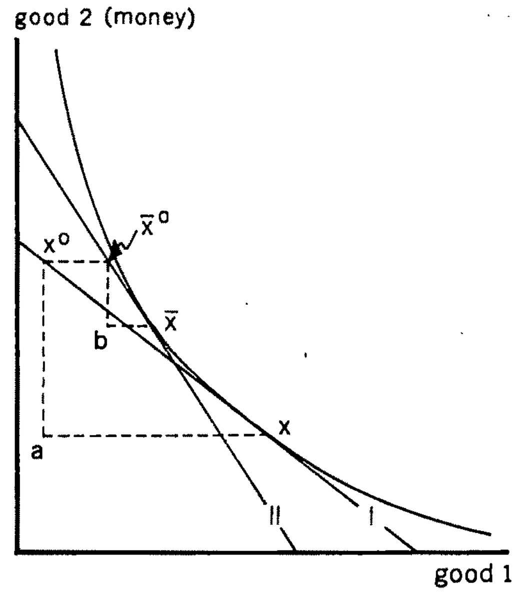



Figure 1. Compensation in Terms of Good 1

It may appear strange, at first glance, that there should be a difference between the effects of compensation in terms of money and compensation in terms of a commodity. For if the consumer is given a few units of commodity $i$ to make up for an increase of $p_i$, he could begin by exchanging those units for money, after which the situation would be conventional. The point is, however, that such an initial exchange may be counted as part of the consumer's transactions. Only if it is so counted—and in Figure 1 it is—will the effect of a compensated price change on expenditures be predictable.

The rest of the paper consists of two sections. In Section I, the result described is

* Assistant professor of economics, State University of New York at Stony Brook. I should like to thank the referee for his helpful comments.

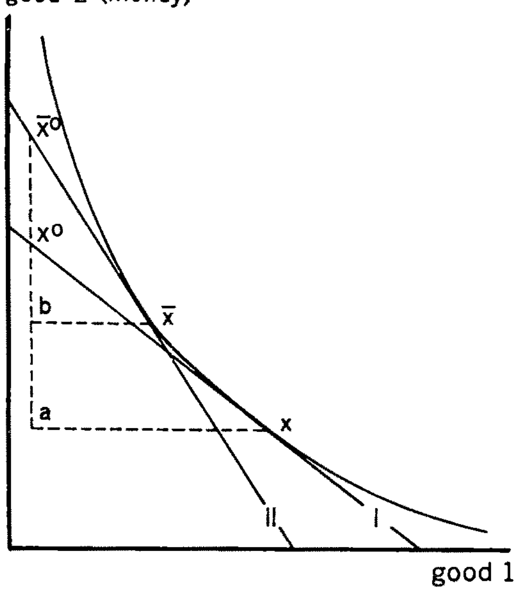



FIGURE 2. COMPENSATION IN TERMS OF GOOD 2

proved for the case of a compensated price change. In Section II, the result is proved for the case of an overcompensated price change, that is, a price change accompanied by such a change of the initial bundle that the old equilibrium bundle becomes reattainable. The overcompensated case is also easy to illustrate geometrically; Figure 1, slightly changed, could serve. Details are omitted.

### I. Effects of a Compensated Price Change

Consider a consumer in a market where $n$ commodities and money are exchanged. Money is regarded as a noncommodity, except when explicitly specified otherwise. Commodity bundles are denoted generically by $\{x_1, \ldots, x_n\}$ or $x$, except for the consumer's initial bundle, written $\{x_1^0, \ldots, x_n^0\}$ or $x^0$. The equilibrium bundle will be represented by $x$; confusion with bundles in general seems unlikely.

In addition to the bundle $x^0$, the consumer brings an amount of money to the market. This amount is $m$, say; the budget constraint is then given by

$$\sum p_j x_j = \sum p_j \overset{0}{x_j} + m$$

or

$$p'x = p'x^0 + m$$

The consumer's behavior is described in terms of a utility function $u(x_1, \ldots, x_n)$ or $u(x)$, maximized, subject to the budget constraint, by the equilibrium bundle $x$. The utility function is taken to be defined over $x_1 > 0, \ldots, x_n > 0$, and to have positive first-order partials, $u_j$, as well as continuous second-order partials, $u_{jk}$, everywhere in its domain of definition. Furthermore, the constrained utility maximum is assumed regular, that is, $d^2u(x)$ is negative. Less stringent conditions would suffice, but those here chosen serve to keep the exposition simple.

Let the price $p_i$ change, all other prices remaining constant. (Throughout, $i$ is arbitrary but fixed.) The price change is assumed to be accompanied by an addition of a possibly negative quantity to $x_i^0$, the initial stock of good $i$. Also, the added quantity is taken to constitute compensation, in the sense that it just enables the consumer to reattain his old utility level. Both the old and the new budget constraints are then tangent to the same indifference surface, so that the envelope theorem applies:

$$0 = \frac{\partial}{\partial p_i}(p'x - p'x^0 - m)$$

$$= x_i - \overset{0}{x_i} - p_i \frac{\partial \overset{0}{x_i}}{\partial p_i}$$

(Since there is no compensation in terms of money, $\partial m / \partial p_i = 0$.) Solving,

$$\partial \overset{0}{x_i} / \partial p_i = (x_i - \overset{0}{x_i}) / p_i \tag{1}$$

If (1) is met, the change of $p_i$ is exactly compensated.

The Slutsky equation is usually derived under the assumption that compensation is made in terms of money, often also under the assumption that the individual's initial commodity bundle is null. Different assumptions are being made here, leading to a different version of the Slutsky equation. This different version, easily found by making minor and obvious changes in any one of the familiar proofs of the Slutsky equation, is

$$\frac{\partial x_j}{\partial p_i} = \left[\frac{\partial x_j}{\partial p_i}\right]_{u=u_0} - \frac{(x_i - \overset{0}{x_i})}{p_i} \cdot \frac{\partial x_j}{\partial x_i^0}$$

In the course of the proof, $\partial x_j/\partial p_i|_{u=u_0}$ is found to be symmetric in $i$ and $j$, and negative if $i=j$. Compensation in terms of good $i$, rather than in terms of money, is thus seen to leave the usual Slutsky properties intact.

This completes the preliminaries. It will now be shown that the price of good $i$ and expenditures for good $i$ change in opposite directions, after compensation. Expenditures for good $i$ are $p_i(x_i - x_i^0)$, and

$$\left[\frac{\partial[p_i(x_i - \overset{0}{x_i})]}{\partial p_i}\right]_{u=u_0} =$$

$$\left[p_i \frac{\partial x_i}{\partial p_i} - p_i \frac{\partial \overset{0}{x_i}}{\partial p_i} + (x_i - \overset{0}{x_i})\right]_{u=u_0}$$

By (1), the right-hand side equals $p_i(\partial x_i/\partial p_i)_{u=u_0}$, which is negative. This proves the assertion.

Compensation in terms of *money* would require that a quantity $\Delta m$ be added to the consumer's initial money holdings, $m$. In the present paper, which takes compensation to be made in terms of a commodity, $\Delta m$ is zero, and $m$ is a constant. The value of this constant has played no role. Our conclusions are therefore valid also if $m=0$, and if $m\equiv 0$. The meaning of $m=0$ is clear. The identity $m\equiv 0$ holds if there exists no money in the economy. It also holds, less obviously, if money is assumed to be a (utility-bearing) commodity. In the latter case, let money be good $r$; then $p_r\equiv 1$, the budget constraint is $p'x=p'x^0$, and the conclusions derived so far are valid for all $i\neq r$. (For $i=r$, the conclusions become meaningless: $p_i$ cannot change if $p_i=p_r\equiv 1$.)

## II. Effects of an Overcompensated Price Change

If a price change is accompanied by an addition to the initial bundle so that the old equilibrium bundle becomes just reattainable, the consumer is said to be overcompensated (see Peter Newman, p. 156). The magnitude of the overcompensation can be objectively determined. Compensation, by contrast, requires that the consumer be able to tell exactly how much of an addition to his initial bundle he needs in order to be exactly as well off as before. Empirically speaking, therefore, overcompensation is preferable to compensation. On the other hand, with overcompensation there does not seem to exist an analogue of the Slutskian symmetry property, and this should perhaps be counted as a weakness of overcompensation. The weakness has its attractive side, however: if a search for the symmetry property is assumed destined to be fruitless, the conditions imposed upon the utility function may be considerably relaxed.

In this section, the utility function is taken to be strictly quasi concave and to have positive first-order partials, everywhere in the positive orthant. As before, utility is maximized subject to the budget constraint by the equilibrium bundle $x$. As before, the budget constraint is given by $p'x=p'x^0+m$.

Unlike before, we make a notational distinction between pre- and postcompensation magnitudes. The latter are marked as such by a bar: $x+dx=\bar{x}$, $x^0+dx^0=\bar{x}^0$, $p+dp=\bar{p}$, and so on. No limits are used in this section. Thus, $d$ represents differences, not differentials; differences of first- or even second-order of smallness are not considered negligible; expressions like $\partial x_j/\partial p_i$ have no meaning.

Let $p_i$ change to $p_i+dp_i=\bar{p}_i$, all other prices remaining constant. Further, let the change be accompanied by an addition $dx_i^0=(x_i-x_i^0)dp_i/\bar{p}_i$ to $x_i^0$, the initial stock of good $i$. It will first be proved that this addition constitutes overcompensation in the sense that it just enables the consumer to reattain his old equilibrium bundle.

The old equilibrium bundle satisfies the old budget constraint, or $p'x=p'x^0+m$. From the definition of $dx_i^0$ as $(x_i-x_i^0)dp_i/\bar{p}_i$ follows that $x_i dp_i=\bar{p}_i dx_i^0+x_i^0 dp_i$; adding this to $p'x=p'x^0+m$ gives $\bar{p}'x=\bar{p}'x^0+m$, showing that $x$ satisfies the new budget constraint, too. This was to be proved.

It is next shown that the Slutskian negative-definiteness property holds: $dp_i dx_i<0$. The individual, enabled by the overcompen-

sation to reattain $x$, chooses $\bar{x}=x+dx$ instead. Apparently, $\bar{x}$ was too costly before, implying

$$p'dx > 0 \tag{2}$$

(A less intuitive proof is the following. By the strict quasi concavity of the utility function, $u_1dx_1+\ldots+u_ndx_n>0$, where the marginal utilities are considered evaluated at the equilibrium bundle. By the first-order condition for a constrained utility maximum, the $u_j$ are positive multiples of the $p_j$. Combining results gives (2).) Since, furthermore, both $\bar{x}$ and $x$ satisfy the new budget constraint, one has $\bar{p}'x=\bar{p}'\bar{x}$, or $\bar{p}'dx=0$. But $\bar{p}'dx=p'dx+dp'dx$, and so, by (2), $dp'dx<0$. By assumption, $p_i$ is the only changing price. We may therefore write, equivalently,

$$dp_idx_i < 0 \tag{3}$$

This was to be shown.

Finally, it will be proved that a change of $p_i$, overcompensated in terms of good $i$, induces a change of expenditures, $p_i(x_i-x_i^0)$, in the opposite direction.

The induced expenditure change is $d[p_i(x_i-x_i^0)]$, which equals $(x_i-x_i^0)dp_i+p_id(x_i-x_i^0)+dp_id(x_i-x_i^0)$, a sum of three terms. The first term equals $\bar{p}_idx_i^0$, by the definition of overcompensation. The other two terms add to $\bar{p}_id(x_i-x_i^0)$. Altogether then, the induced expenditure change is $\bar{p}_idx_i$. Expenditures are thus seen to change in the same direction as $x_i$. By (3), that is the opposite of the direction in which $p_i$ changes. This was to be shown.

Now that an overcompensated price increase has been found to induce an expenditure decrease, it is natural to ask whether the larger of *two* overcompensated price increases will necessarily induce the larger expenditure decrease. The answer is no, as is easy to verify diagrammatically.

The conclusions of this section are, like those of Section I, independent of the value of the consumer's initial money holdings $m$. The conclusions therefore hold if $m=0$ and if $m\equiv 0$; they hold, in particular, when money is regarded as a commodity (other than good $i$).

## REFERENCES

J. R. Hicks, *Value and Capital*, 2nd ed., Oxford 1946.

P. Newman, *The Theory of Exchange*, Englewood Cliffs 1965.

# Liability Rules and Income Distribution in Product Liability

By Koichi Hamada*

The Coase Theorem states that under certain ideal conditions resource allocation is unaffected by liability rules. This result, however, applies only in the absence of transaction or negotiation costs. Moreover, even though resource allocation remains unaffected, the income distribution between the parties is largely influenced by the liability rule. Take the celebrated example of a confectioner and a dentist. The final level of production activity after negotiation between the parties remains the same regardless of which party is liable. But the distribution of income between the confectioner and the dentist depends crucially on which party has the right to the environment.

One of the characteristics of product liability is that producers and consumers are directly or indirectly associated through the price relationship. This very fact gives rise to a case where the conclusion of the Coase Theorem holds true even in the absence of explicit negotiations between the parties. Moreover, under certain conditions, the income distribution in terms of consumers' and producers' surplus could remain the same regardless of the liability rule.

The purpose of this paper is to explore the implication of this price relationship on the product liability. In Section I, I shall analyze the case of product liability where the probability of damage is independent of the level of care taken by producers and consumers. It will be shown in a partial equilibrium framework that not only resource allocation but also income distribution in terms of consumers' and producers' surplus is unaffected by the liability rule, provided that consumers are fully informed of the intensity and the probability of the danger. In Section II, the assumption of the independence of the probability of damage from the care level is dropped. It will be shown that the game theoretic result obtained by John Brown[1] is still valid even if the quantity to be produced and consumed is a variable.

## I. A Partial Equilibrium Analysis of the Product Liability

Let us begin with the analysis of the simplest case where there is a product, say bottled beer, which may cause damage without a safety measure. If the producer spends some additional expenditure on the product, the damage can be prevented. Moreover, we shall assume in this section that the expected value of the damage does not depend on the level of consumers' care nor any other kind of producers' care other than the expenditure on the safety measure.

Suppose the consumers consist of homogeneous individuals and their marginal utility of income can be regarded as constant. Then they are risk neutral, and we are allowed to use the apparatus of partial equilibrium analysis and the concept of consumers' and producers' surplus in a market of a single good.

In Figure 1 let $DD$ be the demand curve for a product on the condition that it is perfectly safe, and let $SS$ be the supply curve of the product without expenditure on the safety measure. Under the validity of the partial equilibrium analysis, the social gain is expressed by the social surplus, that is, the sum of the consumers' and producers' surplus. If the risks from the product, for example, due to the explosion of the bottle, can be neglected, then the maximum social surplus

* Associate professor of economics, University of Tokyo. I am much indebted to Peter Diamond and Iwao Nakajima for their helpful comments on an earlier draft of this paper.

[1] See also Diamond.

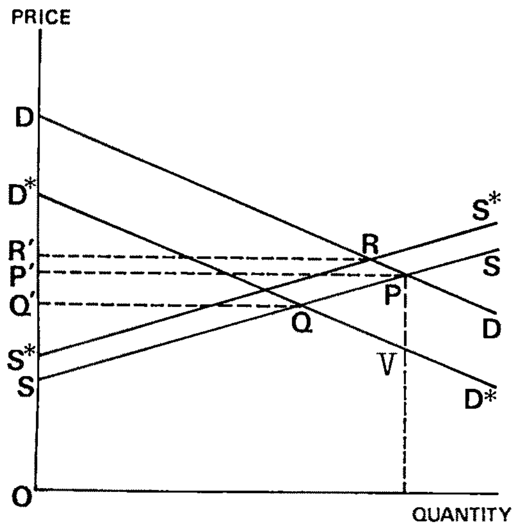



FIGURE 1

is realized as the area of triangle $DPS$ in the competitive market, $OP'$ being the competitive price.

Now we shall turn to the case where there exists danger from the product. If the expected value of the damage of explosion is $DD^*$, the demand curve for the unsafe product becomes as $D^*D^*$ if the consumers know exactly the probability of damage. Also let $SS^*$ be the cost of the safety measure for the product, so that $SS^*$ is the supply curve of the safe product inclusive of the expenditure on the safety measure.

If the producers are liable, they will compare the expected value of the compensation with the cost for implementing the safety measure. Figure 1 is drawn in such a way that $SS^*$ is smaller than $DD^*$. Accordingly, the economic calculation of the producers lets them implement the safety measure in this case because the producers' surplus is larger with the safety measure. Thus equilibrium $R$ is realized, $OR'$ being the price.

If the consumers are liable, that is, if they cannot sue the producers for the damage, it may appear that $Q$ will be realized. However, it is not so. So long as the consumers are perfectly aware of the probability and the magnitude of the damage, it is more profitable for the producers to produce the safe product because the producers' surplus $S^*RR'$ with the safe product is larger than $SQQ'$ with the unsafe product. Thus $R$ is realized in this case as well. Similarly, if $SS^*$ were larger than $DD^*$, the equilibrium corresponding to $Q$ in Figure 1 would be chosen regardless of which party be liable for the damage.

Thus regardless of the liability rule, and even in the absence of explicit negotiations, the competitive market realizes the maximization of the total surplus, provided that the victims are limited to the purchasers of the product, and that the consumers are fully aware of the expected cost of the danger. The direct or indirect price relationship between producers and consumers enables them to engage in implicit negotiations by imputing the burden of risk through the price. Moreover, the distribution between consumers' surplus and producers' surplus remains the same so long as the equilibrium is given by $R$. Thus the liability rule affects neither resource allocation nor income distribution defined in terms of consumers' and producers' surplus. This invariability of income distribution with respect to the liability rule is contrasted with the case of tort in general, where the Coase Theorem gives the invariability of resource allocation but not that of income distribution.

Thus we can state:

PROPOSITION 1: *Neither the resource allocation nor the distribution between the consumers' and the producers' surplus is affected by the liability rule, provided that the consumers are fully aware of the probability and the magnitude of the damage due to the product hazard, and that the victims are limited to the purchasers of the product.*

We should make here, however, the following reservations: First, even though the independence of distribution in terms of producers' and consumers' surplus from the liability rule implies that the *ex ante* expected cost for the damage remains the same for every consumer, it does not preclude the possibility that the *ex post*, that is, after the accident, situation of consumers may differ, depending on the rule. In the absence of a perfect insurance market, the same distribu-

tion in terms of surpluses may mean different social situations. Under caveat venditor, all the consumers pay a higher price, the victims being compensated by the producers; under caveat emptor, the consumers pay a lower price, the unfortunate victims being left uncompensated.

Secondly, the above conclusion applies only when the consumers perfectly foresee the possibility of damage. If for some reason, say, due to the newness of the product, consumers do not foresee the risk, then the demand curve does not shift by the full amount to $DD^*$. Consider the extreme case where $DD$ does not shift at all. Suppose here consumers are not allowed to sue the producers. Since consumers are unaware of the result, there is no incentive for the producers to spend on safety. Accordingly, equilibrium $P$ in Figure 1 is reached with the unsafe product. The producers' surplus equals the area of triangle $SPP'$. Consumers first conceive the consumers' surplus as equal to the area of triangle $DPP'$, but they will realize later that they have to subtract the cost of damage equal to the area of parallelogram $DD^*VP$. Of course, once consumers are fully aware of the danger, $DD$ will shift downward to $D^*D^*$ and the long-run equilibrium will turn out to be $R$. However, it is quite possible that during the transition period consumers lose parallelogram $DD^*VP$ and society loses triangle $QVP$. If the producers are liable, they will shift to the safe product as soon as they realize the danger. Thus even though the dead-weight loss $QVP$ is unavoidable even under caveat venditor so long as the producers are ignorant, the lag to the efficient response could be much shorter in this case.

Thirdly, if the victims are not the purchasers of the product as in the case of any bystanders on the explosion of a bottle, the above reasoning does not apply. Between the third party and the producers, the problem is reduced to that of the effect of the liability rule in usual tort cases such as pollution or automobile accidents. Since it is hardly conceivable that the possible bystanders are able to unite and negotiate for the possible damage, the conditions for the Coase Theorem are not satisfied, the case for caveat venditor could be stated. Of course, if the third party could sue the consumers, then the validity of Proposition 1 would be restored.

Finally, let us notice that the above result is also at least approximately applicable to the case of monopolistic suppliers. In fact, if the expected cost of damage is constant regardless of the quantity produced, $MR$ shifts downward by the amount exactly equal to $DD^*$, leaving the conclusion of Proposition 1 to hold (see Figure 2). Even though the social surplus is not maximized under any liability rule, if the producer monopolizes the market, the amount of commodity produced and the distribution of surplus between consumers and producers are unaffected by the liability rule.

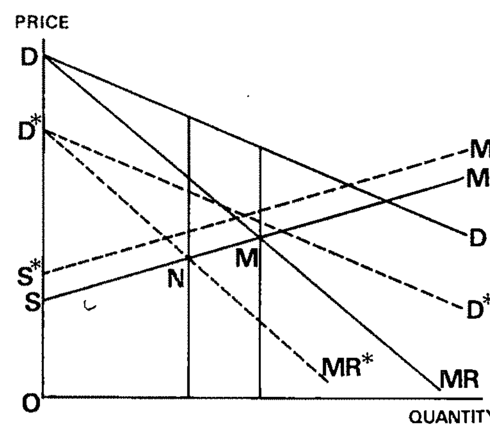



FIGURE 2

## II. The Optimal Level of Care in Product Liability

Now let us relax the assumption that the probability of accidents is independent of the level of care of consumers as well as that of producers. An accident due to a power mower, for example, depends not only on the care taken by its producer but also on the care by its operator. In his illuminating article, Brown (pp. 323 ff) analyzed the interdependence of care by the parties in the framework of a two-person noncooperative game. According to his analysis, the noncooperative solution coincides with the cooperative solution, namely, the Pareto optimal solution under a certain set of rules, pro-

vided that the required levels of care are determined by what he calls the full information incremental standard. The rules are: i) negligence; ii) strict liability with contributory negligence; iii) negligence with contributory negligence; iv) strict liability with dual contributory negligence; v) relative negligence. These rules induce both parties to act in an efficient manner if the required levels of care are set equal to the socially optimal levels.

In Brown's analysis, the question of the optimal combination of care levels is considered under the assumption of a constant level of economic activity. When one wants to apply his analysis of the accident in general to the case of product liability, the question arises whether one can discuss the problem independently of the quantity of the good to be produced and accordingly consumed.[2] Under certain assumptions to be specified below, the optimal liability rules derived by Brown's analysis will be shown to remain optimal even when the quantity produced is a variable.

Let the level of care by the producer and that by the consumer per unit of product be denoted respectively by $X$ and $Y$. They are measured in such a way that the marginal increase in $X$ and $Y$ will incur the constant cost of $W_x$ for the producer and $W_y$ for the consumer per unit of product. The magnitude of the damage per unit of the unsafe product is equal to a constant $A$.[3] The probability of the nonoccurrence of an accident per unit of product is expressed by $\pi(X, Y)$, which is an increasing function of $X$ and $Y$. That is,

$$\pi_x > 0, \qquad \pi_y > 0 \tag{1}$$

Moreover we shall assume

$$\pi_{xx} < 0, \qquad \pi_{yy} < 0, \qquad \pi_{xy} < 0 \tag{2}$$

and[4]

$$\pi_{xx}\pi_{yy} - (\pi_{xy})^2 > 0 \tag{3}$$

Let $L_x(X, Y)$ and $L_y(X, Y)$ be fractions of the damage that must be borne respectively by the producer and by the consumer. Naturally,

$$L_x(X, Y) + L_y(X, Y) = 1 \tag{4}$$

We shall list our assumptions as follows:

Assumption I: The victims of the unsafe product are limited to the buyers of product who know the probability function $\pi(X, Y)$ and the magnitude of damage $A$.

Assumption II: The marginal utility of income for the consumer can be regarded as constant, and the social benefit can be measured by the sum of producers' and consumers' surplus.

Generally speaking, the cost of extra care as well as the probability of accident may depend on the quantity of the product. But for simplicity we shall regard them as independent of the quantity of the product.

Assumption III: The unit costs of care $W_x$, $W_y$, and the probability of prevention of an accident per unit of product purchased $\pi(X, Y)$ are independent of the quantity being produced.

Let $q$ be the quantity of the product to be produced and consumed, and let $p$ be its market price. If we denote by $D(q)$ the (inverse) demand curve for the product in the absence of the possibility of an accident, then the consumers' surplus is expressed as

$$S_y(q, X, Y) = \int_0^q D(q)dq - pq - q\{W_y Y + L_y(X, Y)\,A(1 - \pi(X, Y))\} \tag{5}$$

From Assumption II the consumer is risk neutral, so that we can just deduct from the surplus the expected value of the uncompensated damage. On the other hand, let $C(q)$ be the total cost curve of the product without the cost of implementing the safety measure at all. The producers' surplus is then expressed as

$$S_x(q, X, Y) = pq - C(q) - q\{W_x X + L_x(X, Y)\,A(1 - \pi(X, Y))\} \tag{6}$$

[2] This point is discussed also by W. Y. Oi.

[3] The following qualitative result would still hold even though $A$ is a function of $X$ and $Y$.

[4] Conditions (1) and (2) are explicitly made in Brown, that is, the convexity of the isoquant is also required for the stability of his noncooperative solution. Incidentally, we follow the notation of Brown except for the probability of nonoccurrence of the accident which is denoted here by $\pi$ instead of $P$.

Accordingly, the total social surplus is

$$(7)\quad S_c(q, X, Y) = \int_0^q D(q)dq - C(q) - q\{W_xX + W_yY + A(1-\pi(X, Y))\}$$

The social optimum is achieved when the social surplus is maximized with respect to $X$, $Y$, and $q$. Define the unit cost to the producer $C_x$, the unit cost to the consumer $C_y$, and the unit social cost $C_s$ as:

$$C_x = W_xX + L_x(X, Y)A(1-\pi(X, Y))$$
$$C_y = W_yY + L_y(X, Y)A(1-\pi(X, Y))$$
$$C_s = W_xX + W_yY + A(1-\pi(X, Y))$$

It is easy to see:

*Remark*: If $q$ is given, then the social optimization (= maximization of $S_s$) is equivalent to the minimization of the unit social cost $C_s$. Similarly, if $q$ is given, then the optimization of each party (= maximization $S_x$ or $S_y$) is equivalent respectively to the minimization of the unit cost to the producer or that to the consumer.

Therefore, given $q$, the analysis of the reaction curves made by Brown is valid. The social optimum is achieved by the optimal liability rules derived by Brown, provided that the optimal value of $q$ is realized.

As long as one of the ideal rules is implemented, $X$ and $Y$ are set to the values $X_\Omega$ and $Y_\Omega$ that minimize the unit social cost. Formally, $X_\Omega$ and $Y_\Omega$ are given by $W_x = A\pi_x(X_\Omega, Y_\Omega)$ and $W_y = A\pi_y(X_\Omega, Y_\Omega)$. Here, by Assumption III, $X_\Omega$ and $Y_\Omega$ are independent of $q$, and moreover, identical for every producer and for every consumer. Let $W_xX_\Omega = \hat{\alpha}$, $W_yY_\Omega = \hat{\beta}$, and $\pi(X_\Omega, Y_\Omega) = \hat{\pi}$, then $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\pi}$ are all constant and independent of $q$. Moreover, if we write $L_x(X_\Omega, Y_\Omega) = \overline{L}_x$, $L_y(X_\Omega, Y_\Omega) = \overline{L}_y$, then so long as one of Brown's optimal rules is taken, $\overline{L}_x$ and $\overline{L}_y$ do not depend on $q$ either. Therefore after optimizing with respect to $X$ and $Y$, we can write (5), (6), and (7) as follows:

$$(8)\quad S_y(q) = \int_0^q D(q)dq - pq - q\{\hat{\beta} + \overline{L}_yA(1-\hat{\pi})\}$$

$$(9)\quad S_x(q) = pq - C(q) - q\{\hat{\alpha} + \overline{L}_xA(1-\hat{\pi})\}$$

$$(10)\quad S_s(q) = \int_0^q D(q)dq - C(q) - q\{\hat{\alpha} + \hat{\beta} + A(1-\hat{\pi})\}$$

where $\overline{L}_x + \overline{L}_y = 1$.

Therefore the Pareto-efficient quantity of production $q$ is given by differentiating (10) with respect to $q$,

$$(11)\quad D(q) = C'(q) + \hat{\alpha} + \hat{\beta} + A(1-\hat{\pi})$$

On the other hand, under the perfect competition, consumers maximize (8) with respect to $q$ given $p$ leading to the demand curve:

$$(12)\quad D(q) = p + \hat{\beta} + \overline{L}_yA(1-\hat{\pi})$$

Similarly, supply curve of competitive producers is given by

$$(13)\quad p = C'(q) + \hat{\alpha} + \overline{L}_xA(1-\hat{\pi})$$

The competitive price and quantity are given by a pair $(p, q)$ satisfying (12) and (13). If we take the relation $\overline{L}_x + \overline{L}_y = 1$ into consideration, we can derive (11) from (12) and (13). Thus we have seen that the competitive solution under care levels $X_\Omega$ and $Y_\Omega$ achieves also the Pareto-efficient quantity of production. That is, the care levels $X_\Omega$, $Y_\Omega$, and accordingly the liability rules that are derived as optimum without reference to the level of quantity produced are socially optimal even if the level of economic activity itself is a variable.

Generally speaking $\overline{L}_x$ and $\overline{L}_y$ differ depending on a particular rule among the ideal liability rules. But it is clear from (11) that the values of $\overline{L}_x$ and $\overline{L}_y$ do not affect the value of $q$. Nor do they affect $S_y$ or $S_x$. For by substituting (12) into (8), and (13) into (9), one obtains

$$(14)\quad S_y = \int_0^q D(q)dq - qD(q)$$

$$(15)\quad S_x = qC'(q) - C(q)$$

PROPOSITION 2: *If perfect competition pre-*

*vails under Assumptions I, II, and III, the rules for minimizing the unit social cost also maximize the total social surplus. Neither the level of production (consumption) nor the distribution of surplus between consumers and producers is affected by the liability rule.*

This proposition is a generalization of the analysis of the optimum combination of the levels of care to the case of product liability, and at the same time a generalization of the discussion in the last section that the division of surplus is independent of the liability rule in the context of product liability when the probability of damage and its magnitude are completely foreseen. If the supplier is a monopolist, then the socially optimal quantity is not produced. But the neutrality of the division of surplus between the consumers and the producers with respect to the liability rule still holds.

Let us relax Assumption III and let $W_x$, $W_y$ depend on the amount of production or consumption. In fact, for example, when there are economies of scale in preventing danger—that is, the greater the production, the less costly to produce a safe product—marginal cost of care $W_x$ is a decreasing function of the amount of production or consumption. If the cost of care to an individual producer depends on the amount of his *own* production, and if the cost of care to an individual consumer depends on the amount of his *own* consumption, then the same conclusion as in Proposition 2 will hold true under the following additional assumptions:

Assumption IV: All the producers are identical; all the consumers are identical.

Assumption V: $W_x$ and $W_y$ depend on $q$. But $D(q)-W_y(q)-A(1-\pi(X, Y))$ is a decreasing function of $q$, and $C'(q)+W_x(q)+A(1-\pi(X, Y))$ an increasing function of $q$. Then we can state:

PROPOSITION 3: *If perfect competition prevails under Assumptions I, II, IV, and V, the rules for minimizing the unit social cost also maximize the total social surplus. Neither the level of production (consumption) nor the distribution of surplus between consumers and producers is affected by the liability rule, so long as the society sticks to a single rule.*

Assumption IV is necessary in order to analyze the situation in terms of the representative producer and the representative consumer. Assumption V is required to ensure the second-order condition. The proof goes in a similar manner as that of Proposition 2, utilizing the remark stated there, so that it will be omitted here.[5]

One cannot help feeling uneasy about the above results because the court has to know in advance the socially optimal level of care in order to determine the care standard. In the case of product liability, however, there is a direct or indirect relationship between the consumers and the producers, which enables both parties to engage in implicit negotiations through price. This raises the question whether in some ideal situations the price mechanism itself could achieve the efficient combination of care levels as well as the efficient amount of production without any care standard.

Define the *uninsured price* as

$$p^* = p - L_x A(1 - \pi(X, Y))$$

and the *risk-free price* as

$$p^+ = p + L_y A(1 - \pi(X, Y))$$

The uninsured (supply) price is the price that would prevail if the producer were not liable, and the risk-free (demand) price is the price that would prevail if the consumer were free of any damage. Ultra-considerate or ultra-rational consumers may take the uninsured price as given, and ultra-considerate or ultra-rational producers may take the risk-free price as given in each case instead of the market price. Under these extremely hypothetical assumptions, one can prove the efficiency of price systems without any prescribed care standard, and also the neutrality of the liability rule on the distribution of the

[5] The main difference is that the optimal levels of care $(X_\Omega, Y_\Omega)$ are no longer independent of $q$ but are functions of $q$. But if we notice that $\bar{L}_x$ and $\bar{L}_y$ are independent of $q$ so long as a single rule among the ideal rules is adopted, we can prove the proposition almost in the same way as we proved Proposition 2.

social surplus. Needless to say, this last remark should be interpreted not as the manifestation of the feasibility of the complete market deterrence, but as the indication of the unrealistic and restrictive nature of the assumptions that are required for its feasibility.

REFERENCES

J. P. Brown, "Toward an Economic Theory of Liability," *J. Legal Stud.*, June 1973, *2*, 323–49.

R. H. Coase, "The Problem of Social Costs," *J. Law Econ.*, Oct. 1960, *3*, 1–44.

W. Y. Oi, "The Economic Analysis of Product Safety," *Bell J. Econ.*, Spring 1973, *4*, 3–28.

P. A. Diamond, "Single Activity Accidents," *J. Legal Stud.*, Jan. 1974, *3*, 107–64.

# The Social Cost of Input Distortions: A Comment and a Generalization

By JAMES E. ANDERSON*

Daniel Wisecarver and Richard Schmalensee in two recent papers on the social cost of input market distortions have committed a rather interesting error. They find a different measure of welfare loss due to an input price distortion depending upon whether the measure is in the output or input market. This is incorrect, as intuition surely argues, and the source of the error lies in improper use of the Taylor's series expansion. Correcting the error suggests a worthwhile generalization.

## I. The Error

The error is clearest in the case, considered by both, of price change in a single input used exclusively to produce a single output in competitive partial equilibrium. I first set out the correct analysis. Output $X$ is produced by a degree one homogeneous function of labor $L$ and land $R$, both available in infinitely elastic supply at rates $w_0$ and $r_0$, respectively. Labor is subsidized at rate $\Delta w/w_0$ and the welfare loss is to be computed. Let $a_L \equiv L/X$, $a_R \equiv R/X$. For a degree one homogeneous function, the $a_i$ are functions solely of $w$, $r$: $a_L = a_L(w, r_0)$, $a_R = a_R(w, r_0)$. Then we may use the identity to obtain input demand functions: $L = Xa_L(w, r_0)$, $R = Xa_R(w, r_0)$. In competitive equilibrium with positive production there are no profits, so that $P = a_L w + a_R r_0$, where $P$ is the price of output. Furthermore, by the envelope theorem (unit cost minimization), $w da_L + r da_R = 0$ and thus $\partial P/\partial w = a_L$. All output is sold, so that the output $X = C(P)$, where $C$ is consumption. Measuring the welfare change exclusive of the subsidy payment in the input market yields

$$\Delta S = \int_{w_0}^{w_1} X(P(w))a_L(w)dw$$

Substituting

$(\partial P/\partial w)\cdot dw = a_L(w)dw = dP(w)$, we have

$$\Delta S = \int_{P(w_0)}^{P(w_1)} X(P(w))dP(w)$$

which is the measure of change in the output market. Net welfare changes by $\Delta S + \Delta G$ where[1] $\Delta G = (w_1 - w_0)a_L(w_1)X(P(w_1)) = (w_1 - w_0)L_1$.

How do Wisecarver and Schmalensee obtain different measures? By failing to change $w$ in the output measure to a consistent order of expansion with $w$ in the input market measure. Consider the quadratic approximation in the input market:

$$\begin{aligned}\Delta S &= L(w_0)dw + \frac{1}{2}\frac{\partial L}{\partial w}(w_0)dw^2 \\ &= a_L(w_0)X(P(w_0))dw \\ &\quad + \frac{1}{2}X(P(w_0))\frac{\partial a_L}{\partial w}(w_0)dw^2 \\ &\quad + \frac{1}{2}\frac{dX}{dP}(P(w_0))a_L^2(w_0)dw^2\end{aligned}$$

This translates into the output market as:

$$\begin{aligned}\Delta S &= X(P(w_0))d_2P(w) \\ &\quad + \frac{1}{2}\frac{dX}{dP}(P(w_0))[d_1P(w)]^2\end{aligned}$$

where $d_2P(w)$ denotes the second-order approximation to the change in $P(w)$ and $d_1P(w)$ denotes the first-order approximation to the change in $P(w)$.

* Associate professor of economics, Boston College. I am indebted to J. Huston McCulloch and a referee for helpful comments. I also benefited from conversations with Michael Klass concerning a related paper he coauthored with Alan Deardorff.

[1] Note that $L_1(w_1 - w_0) \neq X(P(w_1))[P(w_1) - P(w_0)]$ in general.

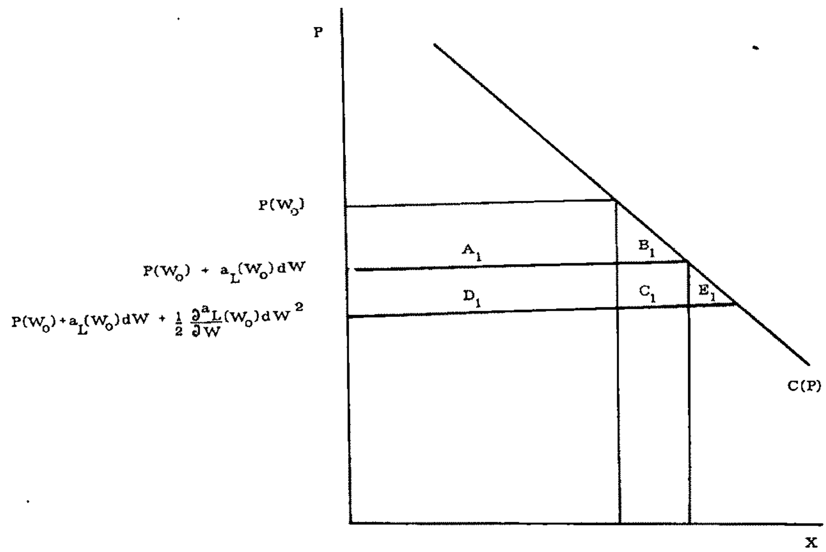



FIGURE 1

In terms of Figure 1, $\Delta S = A_1+B_1+D_1$. Schmalensee argues that $A_1+B_1$ is the correct measure of the change in surplus, while Wisecarver argues for $A_1+B_1+D_1$. Schmalensee's $A_1+B_1$ is not a correct measure under any circumstances. Wisecarver's measure is the true quadratic approximation (in $dw$) in either market. His participation in the error lies in failing to see the source of the problem and accepting Schmalensee's finding that input and output market measures are different. The confusion over consistent expansion can also be seen in higher order terms. Suppose the change in surplus is measured in Figure 1 to the second order in $dP(w)$. The areas $C_1$ and $E_1$ are added to $\Delta S$. They are third and fourth powers of $dw$, but do *not* represent the true third- and fourth-order approximation terms. Measuring $\Delta S$ as $A_1+B_1+C_1+D_1+E_1$ would essentially repeat Schmalensee's error. The Taylor's series approximation always gives the same answer in input and output markets provided it is used to a consistent order of expansion.

A further diagrammatic analysis may be instructive. In Figure 2, land is used as a numéraire ($r_0=1$). As shown $ff'$ is the initial budget line and is tangent to $X_0$ at the initial production point. In terms of land, total cost of $X_0$ is $Of$. The wage decrease causes a shift in cost at constant employment of factors to $Oe$; the difference being $fe$. The value of labor employed in the initial situation is $hf$, and after the tax $he$; the difference again being $fe$. With substitution occurring, the difference rises to $fd$.[2] This demonstrates that in both markets with $dC/dP=0$ $(dX=0)$, the welfare change in either market exclusive of subsidy payment is $\Delta S = fd$. Any consistent degree of approximation of $\Delta S$ must obtain the same measure in either market.

Now consider $X_0$ changing to $X_1$. For simplicity, assume $C$ is linear in $P$. In this case we have:

$$\begin{aligned}\Delta S &= X_0\Delta P + \tfrac{1}{2}(X_1 - X_0)\Delta P \\ &= \tfrac{1}{2}X_0\Delta P + \tfrac{1}{2}X_1\Delta P\end{aligned}$$

Allowing substitution to occur, we obtain:

$$\Delta S = \tfrac{1}{2}fd + \tfrac{1}{2}ab$$

$$\Delta S + \Delta G = \tfrac{1}{2}(fd + ab) - ac$$

This measure is also the net loss in the labor market. If $L(w) = X(P(w))a_L(w)$ is forced to be linear in $w$, a mistake made by both au-

[2] For the labor market, $fd$ is the familiar Hicksian compensating variation measure.

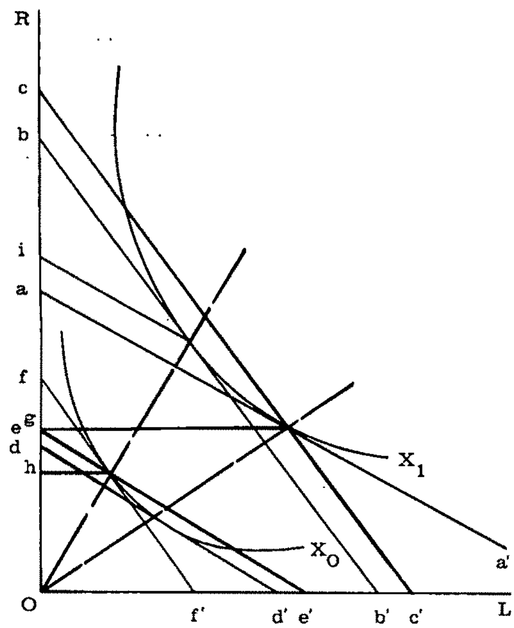



FIGURE 2

thors, we obtain an erroneous different answer.

$$\begin{aligned}\Delta S' &= L_0\Delta w + \tfrac{1}{2}(L_1 - L_0)\Delta w\\ &= \tfrac{1}{2}L_0\Delta w + \tfrac{1}{2}L_1\Delta w\\ &= \tfrac{1}{2}(fe + ac)\\ \Delta S' + \Delta G &= \tfrac{1}{2}(fe + ac) - ac = \tfrac{1}{2}fe - \tfrac{1}{2}ac\\ &= (\Delta S + \Delta G) - \tfrac{1}{2}(ed - bc)\end{aligned}$$

Note that $bc > ed$ in general (given degree one homogeneity and downward sloping final demand). Therefore the input market approximation understates the loss.

The lesson in these mistakes is that quadratic approximations must be used carefully, and perhaps that where theoretical doubts exist, integration should be used.

## II. A Generalization

Studying the correct solution to the error above suggests a generalization. We have seen that: $L(w) = a_L(w)X(P(w)) = X(P(w)) \cdot \partial P/\partial w$. The derivative of the expenditure function (the rate of change of the compensating variation) with respect to the price of labor is precisely the demand for labor. This is quite generally true. Let $S = \sum_i P_iC_i$ be the consumer's expenditure, $(P_1, \ldots, P_n) = P$ be the commodity prices, $(w_1, \ldots, w_m) = w$ be the input prices, and $R_1, \ldots, R_m$ be the factor demands. Make the one-person economy assumption and consider the consequence of changing any factor price.

$$\frac{\partial S}{\partial w_j} = \sum_i \frac{\partial S}{\partial P_i}\frac{\partial P_i}{\partial w_j}$$

With utility held constant,

$$\frac{\partial S}{\partial P_i} = C_i + \sum_k P_k \left.\frac{\partial C_k}{\partial P_i}\right|_{comp} = C_i(P)$$

Then

$$\frac{\partial S}{\partial w_j} = \sum_i C_i \frac{\partial P_i}{\partial w_j} = \sum_i C_i a_{ij} = R_j(w)$$

In measuring the welfare consequences of a factor price change we can integrate either with compensated output demands or with compensated factor demands.[3]

The latter measure involves compensating variations on *factor indifference* curves. These were invented by N. F. Laing for the 2x2 general equilibrium model some years ago. Each such curve gives the locus of factor endowment points which enable utility to be at the given level provided the economy operates with technical efficiency (marginal value products = factor prices and marginal costs = commodity prices). The set of factor indifference curves give (ordinal) utility as a function of endowments. It can be shown that this is a quasi-concave function (Laing has a geometric proof for the 2x2 case).

The geometry may suggest an economy of effort in welfare measurement. Suppose the impact of a vector of factor price changes is to be evaluated. If there are many more factors than goods it may pay to estimate the implied commodity price changes (shifts in unit cost) and then measure the compen-

[3] Note that the symmetry condition holds for the compensated factor demands. This ensures the path independence of the line integral used to evaluate $\Delta S$. Some paths correspond to evaluation over output demands while others correspond to evaluation over input demands.

sating variation by moving a budget plane around an indifference curve in commodity space. On the other hand, with more goods than factors it may pay to aggregate factor demands and measure the compensating variation by moving a budget plane around an indifference curve in factor space.

In practice this discussion suffers from the defects of all surplus analysis—it assumes equilibrium prices are somehow known without solving the general equilibrium problem and it ignores the serious difficulty in inferring aggregate compensating variations from aggregate data.[4] One's reaction to these defects is apparently a matter of taste, but even at a theoretical level there is an appealing convenience to reducing the dimensionality of the compensating variation measure.

[4] For a good discussion, see Donald Richter.

## REFERENCES

N. F. Laing, "A Diagrammatic Approach to General Equilibrium," *Rev. Econ. Stud.*, Feb. 1963, *30*, 43–55.

D. K. Richter, "Games Pythagoreans Play," work. pap., Univ. Rochester 1974.

R. Schmalensee, "Consumer's Surplus and Producer's Goods," *Amer. Econ. Rev.*, Sept. 1971, *61*, 682–87.

D. Wisecarver, "The Social Cost of Input-Market Distortions," *Amer. Econ. Rev.*, June 1974, *64*, 359–71.

# Another Look at the Social Valuation of Input Price Changes

By RICHARD SCHMALENSEE*

In an earlier paper in this *Review*, I considered the Marshallian consumer's surplus valuation of changes in the private and social cost of an intermediate good used by a competitive industry that produced a consumer good under constant returns to scale.[1] I noted that Marshallian measures of the dollar values of such changes could be computed in either the input or final product market and argued that the results would differ unless no substitution were possible in the production of the final good. I concluded that, since the input market can at best reflect the consumer's surplus generated in the final product market, the measure computed in the latter should be used.

Recently, Daniel Wisecarver described this conclusion as "faulty" and "incorrect." Wisecarver was concerned with the social cost of a tax on (or other distortion affecting the market for) an input used in the production of a consumer good; he retained the assumptions of competition and constant returns. He too found that the input and product market valuations would generally differ, but he presented cogent arguments for the superiority of the former measure.

Since Wisecarver did not attempt to exhibit my errors, and since the situations we considered were different, I first suspected that as often happens in economics, we were both right, but about different things. A reexamination of the question, however, did not support this suspicion. We were in fact both wrong and about the same thing. In both situations, the Marshallian valuations computed in the input and product markets are identical. Under competition, either measure can thus be used in practice; the choice between them can properly be made on the basis of convenience.

The following section proves this assertion for a model which contains the two earlier papers as special cases, and it indicates where Wisecarver and I went astray. Section II reexamines a related question considered in my earlier paper: the valuation of input price changes when sellers of the final product have monopoly power.

Let $Q$ be the amount of the final product produced, let $P$ and $MR$ be the corresponding price and marginal revenue, and let $(X_1, \ldots, X_n)$ be the vector of the quantities of the $n$ inputs employed in production.[2] Let $(r_1, r_2, \ldots, r_n)$ be the vector of private and social marginal costs of these inputs. It is assumed throughout the formal analysis that all inputs are in perfectly elastic supply and that $(r_2, \ldots, r_n)$ are fixed, so that we can write $r$ for $r_1$ and $X$ for $X_1$. It is assumed throughout that all the $X_i$ are positive, and the production function for $Q$ is assumed well behaved.

Let $TC(Q, r_1, \ldots, r_n)$ be the minimum private cost of producing output $Q$ with fixed private input prices $(r_1, \ldots, r_n)$, and let $MC(Q, r_1, \ldots, r_n)$ be the corresponding marginal cost function. If the underlying production function is homogeneous of degree $h$, it is easy to show that $TC$ may be written as

$$(1) \quad TC(Q, r_1, \ldots, r_n) = Q^{1/h} G(r_1, \ldots, r_n)$$

where $G$ is concave, nondecreasing, and increasing in $r_1$. Finally, Shephard's lemma implies that the cost-minimizing input demands are given by[3]

$$(2) \quad X_i(Q, r_1, \ldots, r_n) = \partial TC(Q, r_1, \ldots, r_n)/\partial r_i, \quad i = 1, \ldots, n$$

* Associate professor of economics, University of California, San Diego. I would like to thank, with the usual disclaimer, Elizabeth Bailey, George Borts, R. Robert Russell, and Robert Willig.

[1] For an interesting recent discussion of the Marshallian measure, see Willig.

[2] This essay thus relaxes Wisecarver's assumption that $n=2$.

[3] See, for instance, W. E. Diewert, pp. 595–96.

The following two sections consider the social valuation of a change in the social cost of the first input from $r_1$ to $(r_1+\alpha t)$, accompanied by a change in the private cost from $r_1$ to $(r_1+t)$. In Wisecarver's analysis, $\alpha$ was zero, and $t$ was positive, since he was concerned with the impact of taxes or other distortions. My earlier paper was concerned with government actions like research and development projects that would lower both costs equally, so I assumed $\alpha=1$ and $t<0$. A variety of other assumptions are possible. A government investment that lowered the cost of irrigation water and was amortized by pricing the water above marginal cost, for instance, would involve $\alpha>1$ and $t<0$.

Regardless of conditions in the final product market, the Marshallian surplus that would be measured in the market for the first input can be written as

$$(3) \quad S_i(t) = \int_0^{X(t)} \rho(x)dx - (r+\alpha t)[X(t)]$$

where $X(t)$ is the amount of the first input used when its cost is $(r+t)$, and $\rho(x)$ is the inverse demand curve for the first input.[4] The value $S_i(t)$ is simply the usual area under the (input) demand curve, minus the social cost of the (input) quantity actually demanded. The functions $X(t)$ and $\rho(x)$ depend on technology, demand, and behavior in the final product market. Differentiation of (3) yields

$$(4) \quad dS_i(t)/dt = -\alpha X(t) + (1-\alpha)[dX(t)/dt]t$$

The social valuation of the change in $t$ described above that would be computed in the input market is given by

$$(5) \quad V_i(t) = S_i(t) - S_i(0) = \int_0^t [dS_i(\tau)/d\tau]d\tau.$$

[4] That is, $\rho[X(t)]=(r+t)$. If this integral, or the one defined by (7), fails to exist because of the behavior of the inverse demand function near zero, the lower limit of integration may simply be changed to some small positive number, without affecting anything that follows.

For small $t$, the first two terms of the Taylor series expansion about $t=0$ can be used to approximate this quantity:

$$(6) \quad V_i(t) \simeq E_i(t) = t[dS_i(0)/dt] + (t^2/2)[d^2S_i(0)/dt^2]$$

In the final product market, on the other hand, the net Marshallian surplus as a function of $t$ can be written as

$$(7) \quad S_f(t) = \int_0^{Q(t)} P(q)dq - (r+\alpha t)[X(t)] - \sum_{i=2}^{n} r_i X_i(t)$$

This is simply the area under the demand curve minus the social cost of production. Note that factor demands are based on private cost, while inputs are valued at their social costs. The functions $Q(t)$ and $X_1(t) \ldots X_n(t)$ again depend on technology, demand, and behavior in the final product market. An analyst working in the final product market would attempt to measure as accurately as possible the change in $S_f$ caused by a change in $t$. If the Marshallian approach is accepted, it seems obvious that this change is the correct measure of the social value of changes in $t$. Differentiation of (7) yields

$$(8) \quad dS_f(t)/dt = P[dQ(t)/dt] - (r+\alpha t)[dX(t)/dt] - \sum_{i=2}^{n} r_i[dX_i(t)/dt] - \alpha X(t)$$

The quantities $V_f(t)$ and $E_f(t)$ are defined by the obvious analogs of (6) and (7).

## I. Competition

In this section and the next, equilibrium final good output and input demands depend on $t$ only through its effect on the private cost of the first input since production is assumed to minimize private cost, and the private costs of all other inputs and the demand and production functions are assumed fixed. Thus the derivatives of the input demand functions appearing in (4) and (8) are given by

$$(9)\quad dX_i/dt = dX_i/dr = (\partial X_i/\partial Q)(dQ/dr) + (\partial X_i/\partial r),\ i = 1, \ldots, n$$

where the partial derivatives on the right are those of the input demand functions in (2). Similarly, $dQ/dt = dQ/dr$.

Using the superscript $c$ to denote the competitive case, we first prove that $V_i^c(t) = V_f^c(t)$ for any $t$ by showing that $dS_i^c/dt$ and $dS_i^c/dt$ are identically equal for all $t$. If $MP_i$ is the marginal physical product of input $i$, differentiation of the production function for the final product with respect to $r$ yields

$$(10)\quad dQ/dr = \sum_{i=1}^{n} MP_i(dX_i/dr) = \left[(r+t)(dX/dr) + \sum_{i=2}^{n} r_i(dX_i/dr)\right] \Big/ MC$$

where the last equality follows from the private efficiency of production. Under competition $P = MC$ so that substitution for the summation in (8) from (10) yields immediately

$$(11)\quad dS_f^c(t)/dt = -\alpha X + (1-\alpha)(dX/dr)t \equiv dS_i^c(t)/dt$$

from (4), and the proof is complete. Note that it required only private efficiency of production and marginal cost pricing. In particular, it did not require the assumption of linear homogeneity that both Wisecarver and I made in our earlier work. It should also be noted that since it has just been proven that $V_i^c(t) = V_f^c(t)$ for any $t$, it follows that analysis of the two markets will yield precisely the same results even for large $t$.

How then did Wisecarver and I conclude that for small $t$ these measures generally differed by a term in $t^2$? To answer this, let us first differentiate (11) and form the Taylor approximation

$$(12)\quad E_i^c(t) = E_f^c(t) = \alpha Xt - (t^2/2)(1-2\alpha)(dX/dr)$$

where both $X$ and $(dX/dr)$ are evaluated at $t=0$. Under the assumption of linear homogeneity, (1) holds with $h=1$. This implies that $G = MC$,

$$(13)\quad X = Q(\partial G/\partial r) = Q(\partial MC/\partial r)$$

and

$$(14)\quad dX/dr = (dQ/dr)(\partial G/\partial r) + Q(\partial^2 G/\partial r^2) = (dQ/dP)(\partial MC/\partial r)^2 + Q(\partial^2 MC/\partial r^2)$$

using the assumption of marginal cost pricing. Substituting into (12), we have

$$(15)\quad E_i^c(t) = E_f^c(t) = Q(\partial MC/\partial r)t - (1-\alpha)Xt - [(1-2\alpha)/2] \cdot [(dQ/dP)(\partial MC/\partial r)^2]t^2 + [(1-2\alpha)/2] \cdot [Q(\partial^2 MC/\partial r^2)]t^2$$

Both Wisecarver and I used (12) to define $E_i^c$. We used something like (15) to define $E_f^c$, but since we both considered only a first-order approximation to the change in marginal cost, we omitted the final term on the right.[5] Under fixed coefficients, $\partial^2 MC/\partial r^2$ is zero, and we correctly concluded that the two approximations were equal in this case. But when substitution is possible, this quantity is negative, so that its omission from our expressions for $E_f^c$ led us to the conclusion that this quantity was not equal to $E_i^c$ unless substitution was impossible. This conclusion, to repeat, is wrong; it was shown above that under competition, identical measures of the surplus change will be obtained in the affected input market and in the final product market even for substantial changes in input prices and costs.

## II. Monopoly Power

We will use the superscript $m$ to denote the case where the final product is produced by firms with monopoly power. Equation (10) can be used to simplify the expression for $dS^m/dt$ as before, but the substitution

[5] Compare my equation (3) and Wisecarver's equation (4).

$P=MC$ cannot be made. Instead, equations (4), (8), and (10) yield

$$(16)\quad dS_i^m/dt = dS_f^m/dt - (P - MC)(dQ/dr)$$

Assuming the first input is not inferior, $(dQ/dr)$ is negative as long as marginal revenue is declining, and it follows that $dS_i^m/dt$ is greater than $dS_f^m/dt$, the true measure of surplus change. Since an increase in $t$ must serve to lower consumer's surplus, $dS_f^m/dt$ is negative. It then follows that both the social costs of increases in $t$ and the social gain from decreases in this quantity will be underestimated by analysis of the input market.

The basic reason for this bias is simple. Under perfect competition, the net Marshallian surplus $S$ is maximized for given input prices by marginal cost pricing of the final product, so that the partial derivative of $S$ with respect to output is zero. Under imperfect competition, however, this partial derivative is positive since the community as a whole would benefit from increased output. But the welfare effect of *ceteris paribus* output changes is not reflected in input demand functions, and it thus cannot be detected by analysis of input markets.[6]

Things are clearest under the assumption of pure monopoly, which implies $MC=MR$. A monopolist's input demand functions are exactly those of a competitive industry with the same technology and a demand curve equal to the monopolist's marginal revenue schedule. Thus the monopolist's input demands, which depend on marginal revenue products, in effect value changes in final output at $MR$, while the community as a whole values them at $P$. There is simply no way for the $(P-MR)$ difference to show up in analysis of input markets.

One way to salvage input market analysis in this case suggests itself. It might seem that if one knew that the input were sold to a pure monopolist, this knowledge could be used along with additional information to correct the input market valuation of a change in $t$. In fact this can be done for small $t$, but the additional information needed is considerable.

To see this, let $FE_i^m$ and $FE_f^m$ be the first terms in the Taylor approximatious $E_i^m$ and $E_f^m$. To compare these first approximations, first note that under pure monopoly

$$(17)\quad dQ/dr = (dQ/dMR)(dMC/dr) = (dQ/dMR)[(\partial MC/\partial Q)(dQ/dr) + (\partial MC/\partial r)]$$

Second, for any inverse demand function $P=P(Q)$

$$(18)\quad dMR/dQ = 2(dP/dQ) + Q(d^2P/dQ^2)$$

Finally, if the monopoly's production function is homogeneous of degree $h$, equation (1) yields

$$(19)\quad \partial MC/\partial r = X/hQ,$$

$$(20)\quad \partial MC/\partial Q = (1-h)MC/hQ$$

Solving (17) and substituting from (18)–(20), it follows that for a pure monopoly with a homogeneous production function

$$(21)\quad dQ/dr = -XE/PZ$$

where $E$ is the absolute value of the price elasticity of demand

$$(22)\quad Z = h(2-W) + (1-h)(E-1)$$

$$(23)\quad W = (EQ^2/P)(d^2P/dQ^2)$$

It is easy to show that $W$ is less than two if marginal revenue is declining and that the second-order condition for monopoly equilibrium requires $Z$ to be positive.

Equations (4), (6), (16), and (21) along with the first-order conditions for monopoly equilibrium serve to establish the appropriate correction:[7]

$$(24)\quad FE_f^m = FE_i^m[1 + (1/\alpha Z)]$$

where $Z$ is evaluated at $t=0$. First note that if $\alpha$ is zero, no loss is estimated in the input market if terms involving $t^2$ are neglected. In this case, (24) cannot be used and there is no alternative to analysis of the final product

[6] Effects of this sort are briefly discussed in Wisecarver's Section III.

[7] Equation (24) generalizes equation (12) of my earlier paper.

market. Second, even if $\alpha$ is positive, and even in the very special case where $h$ is known to be unity, $W$ must be estimated in order to estimate $Z$. Since $Z$ equals two when final demand is linear under these assumptions, while it equals $(2E-1)/E$ when $E$ is constant, it is clear that one must be quite certain of the form of the final demand function to estimate this quantity with any precision.

Things become even more complicated if a second-order approximation is required. Under the homogeneity assumption, differentiation of (16) and use of (1) and (19)–(21) yields

$$(25)\quad E_f^m = E_i^m - (X/Z)t - (1/2Z)(\partial MC/\partial r)(dQ/dr)t^2 - (hQ/2Z)(\partial^2 MC/\partial r^2)t^2 + (X/2Z^2)(dZ/dr)t^2$$

The second and third terms on the right of this expression are nonzero in general, and they involve the difficult-to-estimate quantity $Z$. The fourth term, which also involves $Z$, is zero if and only if input ratios are fixed by technology. From (22) and (23), the final term vanishes if $E$ is constant along the demand curve, or if demand is linear and $h=1$. In general, from (23), it will involve the third derivative of the inverse demand function.

## REFERENCES

W. E. Diewert, "An Application of the Shephard Duality Theorem: A Generalized Leontief Production Function," *J. Polit. Econ.*, May/June 1971, *79*, 481–507.

R. Schmalensee, "Consumer's Surplus and Producer's Goods," *Amer. Econ. Rev.*, Sept. 1971, *61*, 682–87.

R. D. Willig, "Consumer's Surplus Without Apology," Bell Laboratories econ. dis. pap. 22, Mar. 1975.

D. Wisecarver, "The Social Costs of Input Market Distortions," *Amer. Econ. Rev.*, June 1974, *64*, 359–72.

# IN MEMORIAM

## JOSEPH STANCLIFFE DAVIS

### 1885–1975

Joseph S. Davis, Emeritus Professor of Economic Research and former Director of the Food Research Institute of Stanford University, died in Los Altos, California, on April 23, 1975, at the age of 89. He served as president of the American Economic Association in 1944.

Joseph Davis was born on a small farm in Chester County, Pennsylvania, on November 5, 1885. He graduated from Harvard University summa cum laude in 1908 and was awarded the Doctorate of Philosophy in 1913. He served as a member of the Harvard faculty from 1913 to 1921, with an interruption for service on the Allied Maritime Transport Council in London in 1918–19. In July 1921, he was appointed as one of the three original directors of the Food Research Institute at Stanford University. That Institute was his primary concern until he became emeritus in 1952.

The years from 1921 to 1952 were ones of great achievement in scholarship and institution building. In 1970 Davis recounted how he and the other two directors set out to build an "accurate and increasingly comprehensive body of significant knowledge on world food production, distribution, and consumption." They took as their goals "to analyze, objectively, issues of public policy, and to express the findings clearly and forthrightly, orally and in print." The results of much of this research were embodied in twenty volumes of *Wheat Studies* that were published from 1925 to 1944; they contain twelve special studies and thirty-one survey and review issues of which Davis was author or coauthor, despite the fact that he was away from the university for extended periods in government service. There were also books and very many professional papers. His first book *Essays in the Early History of American Corporation* was published in 1917; his last, *The World Between the Wars*, was published in 1975.

Various of Davis's published essays are landmarks. In August 1949 he presented a paper to the Farm Economic Association at Laramie, Wyoming, entitled "Our Amazing Population Upsurge," that was the opening salvo in his attack on accepted projections of the population of the United States, an attack that ended the interwar belief in a static American population. That paper resulted from his thoroughness in research, his insistence on having the latest information, and his reluctance to trust standing estimates or received doctrine. He had been working for some time on a book on consumption economics, and as he reviewed population projections for the United States and compared them with current population figures he began to sense that something was wrong. As he probed more deeply into the matter, the book was set aside and he concentrated almost exclusively on the population question; it continued to be a matter of special interest into the 1960s.

An article of September 1934, entitled "Agricultural Fundamentalism," is a superb exposition and critique of beliefs then widely held—and still held in some circles—about the place of agriculture in national economies; and it is a model essay, balanced, thoughtful, informed, and lucid.

A third article was published in July 1932 and is particularly apt today. It was entitled "The Specter of Dearth of Food: History's Answer to Sir William Crookes." (It is not, of course, the only article that Davis wrote on this general subject. A particularly cogent statement on the same subject was "Adam Smith and the Human Stomach" in 1954.) Crookes was one of the most eminent scientists of his generation, and in his presidential address to the British Association for the Advancement of Science in 1898 he had forecast a crisis in world food supplies by 1931. Davis carefully examined all of the elements of Crookes's argument and

the quality of his information as a way of getting a better understanding of why the forecast had been so wrong and how better forecasts might be made. He concluded that "political and social factors, not natural or economic limitations, are the principal obstacles to continuous advance in the plane of living."

From his earliest years as an economist, Davis was able to combine scholarly research of the highest quality with the strenuous demands of public service. In 1918–19 he served as an assistant statistician for the American Shipping Mission to London and also as statistician for the Allied Maritime Transport Council. He was a member of the staff of the Dawes Commission on Reparations in 1924, a member of a team studying the Fiscal and Economic Position of Mexico in 1928, chief economist of the Federal Farm Board in 1929 to 1931, a member of the Food and Nutrition Board of the National Research Council from 1940 to 1945, and of the Institute of Pacific Relations from 1941 to 1947. In 1955, three years after his official retirement, he was called to serve for nearly four years as a member of President Eisenhower's Council of Economic Advisors.

The years were also not without their honors. In addition to being president of this association, Davis was elected president of both the American Statistical Association and the American Farm Economic Association in 1936, and in 1957 he was made one of the first Fellows of the American Farm Economic Association. But he cherished just as much the "Distinguished Alumnus Award" he received in 1973 from Westchester (Pa.) State College, where he graduated in 1902.

Davis used to say that he felt more at home in research than in teaching, and that he had no real gift for classroom teaching. He did in fact have a great gift for teaching, by a word here, a sharp question there, a turn of phrase, or an aphorism. And many young scholars found his probing interest in their research the finest reward they could wish for and the greatest possible stimulus to their best efforts. At the same time, he rarely engaged in small talk, and those who did not know him were sometimes taken aback by the abruptness with which business conversations began and ended.

Joseph Davis was a teacher as well as a devoted seeker of knowledge and a loyal servant of his society. As he taught himself, he taught others, and his persistent emphasis on "expressing findings clearly and forthrightly" was an essential element in this teaching. He never lost his desire to know, and he was indomitable in overcoming ignorance. In a talk to a Stanford undergraduate group in 1961 he said "Life has been far tougher, but much richer, than I dimly anticipated when I was twenty." He would surely have said the same thing fourteen years later.

# NOTES

Economists who are *strongly* oriented toward the humanities, who use humanistic methods in their research, and who will be participating in meetings abroad that are concerned with the humanistic aspects of their discipline are eligible to apply for travel grants of the American Council of Learned Societies. Specifically, economists may be eligible if (a) they deal with the history of economic thought or economic history, and (b) if their approach is qualitative and descriptive rather than quantitative and statistical. Conferences dealing with the establishment of social policy or legislation are generally ineligible. The deadlines for applications to be received in the office of the American Economic Association are: for meetings scheduled between June and September, February 15; for meetings scheduled between October and January, June 15; for meetings scheduled between February and May, October 15. Application forms may be obtained from C. Elton Hinshaw, Secretary, American Economic Association, 1313 21st Avenue South, Nashville, Tennessee 37212.

---

Nicolaus and Associates have been appointed the official travel coordinator for the Atlantic City meetings of Allied Social Science Associations. Group air fares may be arranged from your city and for groups as small as fifteen. These fares usually run about 25 percent below regular coach fare. Some examples using current fares subject to change are:

| Round Trip: | Coach | Group |
|---|---|---|
| New York-Atlantic City | $ 61.46 | $ 41.46 |
| Detroit-Atlantic City | 161.46 | 87.46 |
| San Francisco-Atlantic City | 389.46 | 310.46 |

Flights will be arranged for other cities if enough individuals show interest. Air reservation forms and general information on flights and costs will be mailed to you in April with the preregistration material. In the meantime if you have any questions on group flights, please call Nicolaus and Associates collect at 800+433-2897, or write them at 1201 N. Watson Road, Suite 118, Arlington, Texas 76011.

---

The Research Program on Population and Development Policy, sponsored jointly by the Rockefeller and Ford Foundations, is interested in receiving proposals that examine empirical relationships between development processes and population trends; reciprocal effects of development and population policies; and means by which development processes and population trends might be modified to achieve improvements in human welfare. The submission deadline is July 1, 1976 and awards will be announced in December. For further information outlining application procedures, write The Rockefeller-Ford Foundations' Research Program on Population and Development Policy, The Rockefeller Foundation, 1133 Avenue of the Americas, New York, NY 10036.

---

*Change Magazine*, under a grant from the Fund for the Improvement of Post Secondary Education, has launched a program of undergraduate teaching recognition. The purpose of the program is to identify, with the cooperation of the disciplinary associations, examples of the best efforts to improve undergraduate teaching. Faculty members so identified will receive recognition through publication of a detailed description of their work in a series of reports on teaching. In order to participate, faculty members must share their experiences, through written material and in person, with other teachers, students, and outside evaluators. The experience should be adaptable to other institutions and, if possible, to other disciplines. A complete announcement of this program, together with the criteria to be used in choosing the courses recognized, was published in *Change Magazine*, November 1975. Faculty members interested in submitting an application for consideration should send a brief description of their work to C. Elton Hinshaw, Secretary, American Economic Association, 1313 21st Avenue South, Nashville, Tennessee 37212.

---

*Fulbright-Hays Program for Senior American and Foreign Scholars*

Applications will be accepted this spring for more than 500 university lecturing and advanced research awards during 1977–78 in over 75 countries under the senior Fulbright-Hays program, the Council for International Exchange of Scholars announced recently. Specialists in economics and business administration who are U.S. citizens and have a doctorate or college teaching experience are invited to indicate their interest in an award by completing a simple registration form, available on request from the Council, Eleven Dupont Circle, Washington, D.C. 20036. In April 1976, each registrant will be sent an announcement of opportunities under the 1977–78 program. Applications for 1976–77 are at present under review, but some awards remain open to application. Inquiries about remaining openings are welcomed by the Council.

Each year Fulbright-Hays agencies abroad forward to the Council applications of senior foreign scholars who are interested in remunerative appointments for lecturing and postdoctoral research at American universities or colleges for temporary periods. Approximately 500 foreign scholars are awarded grants to come to the United States after arrangements are made for lecturing or research assignments. Colleges or universities interested in having a foreign Fulbright-Hays scholar on campus during 1976–77 should write to the Council as soon as possible. A directory of foreign Fulbright lecturers and research scholars in the United States is available on request to the Council.

The Fifth World Congress of the International Economic Association, "Economic Growth and Resources," will be held in the Tokyo Prince Hotel, from August 29 to September 3, 1977. The Program Committee is under the Chairmanship of Edmond Malinvaud (France), president of the International Economic Association, with the collaboration of Shigeto Tsuru (Japan), Mekki Abbas (Sudan), Moses Abramovitz (USA), S. Chakravarty (India), T. Khachaturov (USSR), Assar Lindbeck (Sweden), R. Matthews (UK), D. Sadowski (Poland), Robert Solow (USA), O. Sunkel (Chile), H. Uzawa (Japan), C. von Weizsäcker (FRG).

Arrangements are being made to provide special travel and accommodation for participants, including, when desired, pre- or post-Congress tours. Participants will receive periodic information on the progress of its preparation on request to the Local Organizing Committee (Chairman, Shigeto Tsuru), Tokei Kenkyu Kai, 1-10 Shinbashi 4-chome, Minato-ku, Tokyo 105, Japan.

Suggested contributions to the scientific work of the Congress should deal with the general topic of "Economic Growth and Resources," broadly understood, and be submitted to the Members of the Program Committee or sent to the Secretariat of the International Economic Association in Paris, 54 Boulevard Raspail (bureau 428), 75270 Paris Cedex 06, for transmission to the Committee.

---

The Administration on Aging of the Department of Health, Education, and Welfare announces continuation of its Dissertation Research Grant Program for 1976. A limited number of stipends of $5,000 each will be awarded to graduate students at the Ph.D. level in social science who agree to devote one year full time to the research and writing of a dissertation that focuses on the elderly or the aging process. Students in social gerontology, sociology, social work, economics, psychology, political science, public administration, and other relevant social sciences are eligible. The application deadline is March 31, 1976. Copies of the Guidelines can be obtained by writing Dissertation Fellowship Program, Division of Research and Analysis, Department of Health, Education, and Welfare; Office of Research, Demonstrations, and Manpower Resources, Administration on Aging, 400 Sixth Street, S.W., Washington, D.C. 20201.

---

The Conference on Social Sciences in Health is seeking contributed papers for part of its program at the annual meetings of the American Public Health Association in Miami Beach, Florida, on October 17–21, 1976. Interested persons should submit an abstract of up to 200 words as promptly as possible to Irving Leveson, Senior Professional Staff, Hudson Institute, Quaker Ridge Road, Croton-on-Hudson, NY 10520.

---

Omicron Delta Epsilon, the International Honor Society in Economics, invites the submission of entries for the eighth year of the Irving Fisher Graduate Monograph and Frank W. Taussig Award Undergraduate Competitions. The Fisher Award consists of $1,000 and publication as a book by Princeton University Press, subject to approval of its editorial board. In addition, the winner will be invited to submit a paper based on the winning entry to the *American Economic Review*. The recommendations of the Final Selection Board of the Competition will be considered by the *Review* in the refereeing process. All finalists will be invited to submit a paper for publication in *The American Economist*. The Taussig Award consists of $100 and publication in *The American Economist*. Entries for the Fisher Award should be submitted to Departmental Selection Committees by January 1, 1976 and entries for the Taussig Award by May 15, 1976. They will be judged by the International Editorial Board and finalists by the Final Selection Board, consisting of Professors William J. Baumol, Frank H. Hahn, Robert M. Solow, Arnold Zellner, and Egon Neuberger (editor). For more information, write Professor E. Neuberger, Editor, Economic Research Bureau, State University of New York, Stony Brook, NY 11794.

---

The Faculty Exchange Center, founded and administered by faculty members, helps to arrange college and university faculty exchanges within the United States and overseas where the language of instruction is English. For more information, write to Faculty Exchange Center, P.O. Box 1091, Lancaster, Pennsylvania 17604.

---

Papers are solicited for the eighth Amos Tuck School Seminar on Problems of Regulation and Public Utilities, sponsored by the American Telephone and Telegraph Company, to be held August 1–5, 1976 at Dartmouth College. The seminar will be attended by thirty conferees chosen from academic institutions throughout the country. It consists of seven or eight sessions over a four-day period. The sessions provide a forum for the presentation and discussion of new ideas and innovative work in analyzing regulated firms, regulatory processes, and problems of public welfare. Those with an interest in this area are invited to suggest names of possible speakers or to apply themselves. The aim of the seminar is to introduce the research both of young scholars beginning their careers and of scholars who have only recently turned their attention to public utility problems. Therefore, suggestions of speakers must be limited to scholars in these categories. Those chosen to present papers will receive support for research and writing, in addition to travel and living expenses at the seminar. This is not a request for conferee applications. Application forms and further information will be mailed to institutions in March 1976. The directors of the seminar are Professor Duncan M. Holthausen, Jr., and Professor Dennis E. Logue. Please direct all suggestions and correspondence to Directors, Seminar on Problems of Regulation and Public Utilities, The Amos Tuck School of Business Administration, Dartmouth College, Hanover, New Hampshire 03755.

Papers are being accepted for presentation at a conference, "Major International Policy Issues," to be held at the Malibu (California) Campus of Pepperdine University in the fall of 1976. The papers are expected to be policy oriented and to explicate the various dimensions of and alternate policies and consequences related to such major international issues as resource shortages, indexation, detente and disarmament, worldwide stagflation, comparative productivity, comparative quality of life, new economic order, food and population, evolution of economic systems, managed growth vs. limits to growth, trade and aid, multinational corporations, transfer of technologies, international oligopolies, and world monetary systems. Maximum size of papers is 40 typewritten double-spaced pages. Proceedings will be published. Interested scholars should submit preliminary outlines to Professor Nake M. Kamrany, Massachusetts Institute of Technology, Bldg. 20B-221A, Cambridge, MA 02139.

---

The tenth annual meeting of the Pacific Northwest Regional Economic Conference will be held May 6–8, 1976 in Victoria, British Columbia. Academic, business, and government economists concerned with the U.S. and Canadian northwest economies are invited to participate. Persons interested in presenting papers on any subject relevant to the northwest economy are invited to submit abstracts to the program chairman, Professor G. R. Walter, Department of Economics, University of Victoria, Victoria, B.C., Canada, V8W 2Y2.

---

### *Deaths*

Ralph E. Burgess III, international economist, U.S. Treasury Department, Jan. 3, 1975.

Carroll L. Christenson, department of economics, Indiana University, Bloomington, Oct. 11, 1975.

Edward Everett Hale, emeritus professor of economics, University of Texas, Austin, Feb. 3, 1975.

Samuel M. Levin, emeritus professor, Wayne State University, Oct. 2, 1975.

### *Retirements*

J. C. D. Blaine, University of North Carolina, Chapel Hill, June 30, 1975.

Robert W. Bradbury, professor of economics, University of Florida, June 1975.

Elmo L. Jackson, professor of economics, University of Florida, June 1975.

Donald Marschner, professor of business administration, Whittemore School of Business and Economics, University of New Hampshire, July 1, 1975.

Clarence E. Philbrook, University of North Carolina, Chapel Hill, May 31, 1975.

### *Visiting Foreign Scholars*

Emil-Maria Claassen, University of Paris: visiting professor of economics, University of Florida, Jan.–Mar. 1976.

Ehud Kalai, Tel Aviv University: department of managerial sciences, Northwestern University, 1975–76.

Jean Parent, Université de Paris I: visiting scholar, department of economics, Université de Moncton, Sept. 1975.

### *Promotions*

Lewis M. Abernathy: professor of economics, North Texas State University, Sept. 1, 1975.

Byron B. Brown, Jr.: professor of economics, Southern Oregon State College, Sept. 1, 1974.

Steven T. Call: assistant professor, department of economics and Center for Community Leadership Development, University Extension, University of Wisconsin, Milwaukee, July 1, 1975.

Sidney L. Carroll: associate professor, department of economics, University of Tennessee, Knoxville, Sept. 1, 1975.

David A. Denslow: associate professor of economics, University of Florida, Mar. 1975.

Joseph S. De Salvo: professor, department of economics, University of Wisconsin, Milwaukee, July 1975.

Richard J. Gelson: chief, Financial Statistics Division, Federal Reserve Bank of New York, Oct. 16, 1975.

Morley Gunderson: associate professor, University of Toronto, July 1975.

John S. Hill: senior economist, Open Market Operations and Treasury Issues Function, Federal Reserve Bank of New York, Sept. 18, 1975.

Ivan Iskroff: special assistant, foreign department, Federal Reserve Bank of New York, Oct. 2, 1975.

Feng-Yao Lee: professor, department of economics, University of Tennessee, Knoxville.

Fred J. Levin: manager, domestic research department, Federal Reserve Bank of New York, Sept. 18, 1975.

Steven An-Yhi Lin: professor of economics, Southern Illinois University, July 1975.

Milo F. McCabe: associate professor, department of economics, University of South Dakota, Aug. 1975.

Ann-Marie Meulendyke: chief, Securities Analysis Division, Federal Reserve Bank of New York, Oct. 16, 1975.

Peggy B. Musgrave: professor of economics, Northeastern University, July 1, 1975.

C. Louise Nelson: professor of economics, Davidson College, July 1975.

Albert W. Niemi, Jr.: professor, department of economics, University of Georgia, July 1975.

Paul H. Rubin: associate professor, department of economics, University of Georgia, Sept. 1975.

Pawan K. Sawhney: assistant professor of economics, Northeastern University, July 1, 1975.

Richard M. Scheffler: associate professor of economics, University of North Carolina, Chapel Hill, July 1, 1975.

Stephen L. Shapiro: associate professor of economics, University of North Florida, July 1975.

Dennis E. Smallwood: associate professor of economics, University of California, San Diego, July 1, 1975.

Sheila L. Tschinkel: adviser, Open Market Operation

and Treasury Issues Function, Federal Reserve Bank of New York, Sept. 18, 1975.

William G. Tyler: associate professor of economics, University of Florida, Mar. 1975.

Mira Wilkins: professor of economics, Florida International University, Sept. 1975.

### *Administrative Appointments*

David F. Baron: chairman, department of managerial economics and decision sciences, Northwestern University.

Joel Bergsman, The Urban Institute: chief, community Development and Housing, Congressional Budget Office, Sept. 15, 1975.

Emery N. Castle, Oregon State University: senior fellow and vice president, Resources for the Future, Inc., Jan. 1, 1976.

Arnold Collery, Council on Wage and Price Stability: professor of economics and acting dean of the faculty, Amherst College, July 1, 1975.

Stephen L. Fink: associate dean, Whittemore School of Business and Economics, University of New Hampshire, Sept. 1, 1975.

Joseph W. Ford: chairman, department of economics, St. Francis College, Sept. 1975.

Richard T. Froyen: chairman, graduate studies in economics, University of North Carolina, Chapel Hill.

Stuart I. Greenbaum: chairman, economics department, College of Business and Economics, University of Kentucky, Aug. 1, 1975.

Guy C. Grenier, Florida Department of Agriculture and Consumer Services: assistant dean, Jones College, Aug. 1975.

David R. Kamerschen: head, department of economics, University of Georgia, July 1, 1975.

Michael J. Lavell: dean, School of Business, John Carroll University, Aug. 15, 1975.

Ahmad A. Murad: chairman, department of economics, Eastern Illinois University, Aug. 1975.

James L. Murphy: chairman, department of economics, University of North Carolina, Chapel Hill.

Frank J. Navratil: chairman, department of economics, John Carroll University, July 1, 1975.

Albert W. Niemi, Jr.: acting director of research, College of Business Administration, University of Georgia, Sept. 1975.

Arnold W. Sametz: director, Salomon Center for the Study of Financial Institutions, Graduate School of Business Administration, New York University, Sept. 1, 1975.

Sam H. Schurr, Electric Power Research Institute: senior fellow, Resources for the Future, Inc., Jan. 1, 1976.

Calvin D. Siebert: chairman, department of economics, University of Iowa.

Kenneth Smith, University of Wisconsin: director, hospital and health services management program, Graduate School of Management, Northwestern University.

A. H. Studenmund: chairman, department of economics, Occidental College, Sept. 1975.

D. Babatunde Thomas: chairperson, department of economics, Florida International University, Mar. 1975.

William G. Tyler: acting chairman, department of economics, University of Florida, Sept. 1975.

Walter A. Verdon: director, Cleveland Center for Economic Education, John Carroll University, July 1975.

### *Appointments*

James D. Adams: instructor, department of economics, Iowa State University, Sept. 1, 1975.

Neil O. Alper: research assistant professor, department of economics and Center for Business and Economic Research, University of Tennessee, Knoxville.

Arthur T. Andersen: visiting professor of economics, College of William and Mary, Sept. 1975.

Robert H. Andrew: assistant professor, department of economics, University of Georgia, Sept. 1975.

Robert C. Baesemann, Washington University: assistant professor, department of managerial economics and decision sciences, and the Transportation Center, Northwestern University.

Kenneth Boulding: visiting professor, department of economics, Wellesley College, fall 1975.

Helen Cammarata, University of Maryland: instructor in economics, Wellesley College, fall 1975.

Anne P. Carter, Brandeis University: visiting professor, department of economics, New York University, Sept. 1975.

George Chen: instructor of economics, Northeastern University, Sept. 29, 1975.

W. Helena Choynacka: visiting assistant professor, Ohio State University.

William K. Clarritt, Rutgers University: assistant professor of economics and business, William Paterson College, Sept. 1, 1975.

Bryan C. Conley: assistant professor of business economics, University of Southern California, Sept. 1, 1975.

Jacques Cremer: instructor of economics, Northeastern University, Sept. 29, 1975.

Andrew Daughety, U.S. Air Force: assistant professor, department of managerial economics and decision sciences, and the Transportation Center, Northwestern University.

David B. H. Denoon, Massachusetts Institute of Technology: assistant professor, department of economics, New York University, Sept. 1975.

Arthur G. Dobbelaere, Jr., Notre Dame: assistant professor of economics, Loyola University, Sept. 1, 1975.

Arthur L. Dolinsky, University of Pennsylvania: instructor of economics and business, William Peterson College, Sept. 1, 1975.

Jacob S. Dreyer: assistant professor, department of economics, New York University, Sept. 1975.

James E. Duffy: visiting instructor in economics, John Carroll University, Sept. 1975.

Walter W. Ebanks: research associate, National Bureau of Economic Research, Oct. 1, 1975.

James W. Eiler: instructor, department of economics, Iowa State University, Sept. 1, 1975.

Robert F. Engle, Massachusetts Institute of Tech-

nology: associate professor of economics, University of California, San Diego, July 1, 1975.

Henry H. Fishkind: assistant professor of economics, University of Florida, June 1975.

James Fortson, University of Georgia: regional economist, 6th National Bank Region, Atlanta.

Daniel J. Gallagher, University of Cincinnati: assistant professor of economics and commerce, Niagara University, Sept. 1, 1975.

Roy J. Gardner: assistant professor, department of economics, Iowa State University, Sept. 1, 1975.

Hugh B. Garnett, Southern Methodist University: assistant professor, department of economics, North Texas State University, Sept. 1975.

Gilbert R. Ghez, University of Chicago: associate professor of economics, University of Illinois at Chicago Circle, Sept. 1975.

Michael L. Goetz, Saint Mary's College: assistant professor of economics, University of Tulsa, Sept. 1975.

Stephen E. Guisinger: research associate, National Bureau of Economic Research, Oct. 1, 1975.

Richard D. Gustely: research assistant professor, department of economics and Center for Business and Economic Research, University of Tennessee, Knoxville.

John W. Hagens, Cornell University: assistant professor of economics, Colby College, Sept. 1, 1975.

Russell I. Haley: associate professor of business administration, Whittemore School of Business and Economics, University of New Hampshire, Sept. 1, 1975.

Robert Halvorsen: research associate, National Bureau of Economic Research, Sept. 1, 1975.

Robert Hankin: instructor of economics, Northeastern University, Sept. 29, 1975.

Hiroaki Hayakawa: assistant professor, department of economics, University of Georgia, Sept. 1975.

J. Stephen Henderson, University of Wisconsin: visiting assistant professor, Ohio State University.

Marc W. Herold: instructor in economics, Whittemore School of Business and Economics, University of New Hampshire, Sept. 1, 1975.

R. Carter Hill: assistant professor, department of economics, University of Georgia, Sept. 1975.

John Hiller: assistant professor of economics, University of Wisconsin, Milwaukee, Sept. 2, 1975.

Peter N. Hopcraft: research associate, National Bureau of Economic Research, Oct. 1, 1975.

Thomas D. Hopkins, Irwin Management Company: economist, Council on Wage and Price Stability, Apr. 1975.

James H. Hugon, Portland State University: regional economist, 13th National Bank Region, Portland.

Dale W. Janowsky: instructor of economics, Niagara University, Sept. 1, 1975.

Stanley R. Johnson: professor, department of economics, University of Georgia, Sept. 1975.

James E. Jonish, Texas Tech University: visiting associate professor, department of economics, Dartmouth College, Sept. 1975.

John Joseph, Jr., Maine State Department of Commerce and Industry: instructor of economics, Colby College, Sept. 1, 1975.

John P. Judd: economist, Money and Finance Division, Federal Reserve Bank of New York, Oct. 16, 1975.

Nicholas Karatjas: assistant professor of economics, Syracuse University, Sept. 1, 1975.

Lewis Karstensson, Ohio University: assistant professor of economics and associate director, Center for Economic Education, and Manpower and Industrial Relations Institute, North Texas State University, Sept. 1, 1975.

Lawrence W. Kenny: assistant professor of economics, University of Florida, June 1975.

Kenneth W. Koford, University of California, Los Angeles: instructor, department of economics, Occidental College, July 1975.

Claudio Kraus: instructor of economics, Northeastern University, Sept. 29, 1975.

Donald Leet, University of California, Fresno: assistant professor of economics, Colby College, Sept. 1, 1975.

Philip Levine, Columbia University: assistant professor of economics and business, William Paterson College, Sept. 1, 1975.

An-Loh Lin: economist, Business Conditions Division, Federal Reserve Bank of New York, Oct. 16, 1975.

Charles W. Lindsey: instructor, department of economics, Trinity College, Sept. 1, 1975.

Joseph A. Lisciandro: instructor, department of economics and business, Slippery Rock State College, Sept. 1975.

Royce A. Lorentz: assistant professor, department of economics and business, Slippery Rock State College, Sept. 1975.

Lawrence K. Lynch: visiting associate professor of economics, University of Kentucky, 1975–76.

Thomas S. McCaleb: assistant professor of economics, University of Kansas, Aug. 1975.

John F. McDonald: associate professor of economics, University of Illinois at Chicago Circle, Sept. 1975.

Robert McLean: instructor of economics, University of Wisconsin, Milwaukee, Sept. 2, 1975.

G. S. Maddala: professor of economics, University of Florida, Sept. 1975.

Stanley Mahlahla: instructor of economics and business, William Paterson College, Sept. 1975.

Rita M. Maldonado-Bear: associate professor of finance and economics, Graduate School of Business Administration, New York University, Sept. 1, 1975.

Lewis Mandell, Southern Methodist University: special economic consultant, Comptroller of the Currency, Washington, D.C.

John T. Masten, University of Georgia: visiting professor of economics, University of Kentucky, 1975–76.

Laurence H. Meyer, Washington University, St. Louis: adjunct assistant professor, department of economics, New York University, Sept. 1975.

John R. Milo, New York University: assistant professor of economics and business, William Paterson College, Sept. 1, 1975.

Michael A. Mogavero, University of Connecticut: assistant professor of economics, Niagara University, Sept. 1, 1975.

Andrew Moody: instructor of economics, Northeastern University, Sept. 29, 1975.

William S. Mounts: temporary instructor, department of economics, University of Georgia, Sept. 1975.

Charles Mueller: instructor of economics, Northeastern University, Sept. 29, 1975.

Neil B. Murphy, University of Maine: regional economist, 1st National Bank Region, Boston.

Joel Naroff: instructor of economics, Northeastern University, Sept. 29, 1975.

Kenneth J. Nicol: assistant professor, department of economics, and Center for Agricultural and Rural Development, Iowa State University, July 1, 1975.

Joseph Pelzman: instructor of economics, Northeastern University, Sept. 29, 1975.

Doyle V. Peterson: research associate, department of economics, Iowa State University, Sept. 1, 1975.

Robert L. Pirog, Columbia University: instructor of economics and business, William Paterson College, Sept. 1, 1975.

Dean R. Plager: instructor, business administration, Whittemore School of Business and Economics, University of New Hampshire, Sept. 1975.

Leonard Presby, New York University: assistant professor of economics and business, William Paterson College, Sept. 1, 1975.

David H. Pyle, University of California, Berkeley: regional economist, 14th National Bank Region, San Francisco.

Gregory Richardson, Cornell University: professor of economics, University of Kentucky, July 1, 1975.

Steven M. Rock, Northwestern University: visiting assistant professor of economics, University of Illinois at Chicago Circle, Sept. 1975.

Roy T. Savoian, University of California, Santa Barbara: assistant professor of economics, University of Tulsa, Sept. 1975.

Alan M. Schlottman: assistant professor of economics, University of Tennessee, Knoxville.

Andrew Schotter, Syracuse University: assistant professor, department of economics, New York University, Sept. 1975.

David L. Scott, Florida Southern College: associate professor of accounting and finance, Valdosta State College, Sept. 1, 1975.

Richard Shick, Saint Bonaventure University: regional economist, 4th National Bank Region, Cleveland.

Constantino Sifniotis: instructor of economics, Northeastern University, Sept. 29, 1975.

Kenneth Smith, University of Wisconsin: professor, department of managerial economics and decision sciences, Northwestern University.

John F. Stewart, University of Wisconsin: assistant professor of economics, University of North Carolina, Chapel Hill, July 1, 1975.

Roger Stover, University of Minnesota: regional economist, 9th National Bank Region, Minneapolis.

Stephen H. Strand: assistant professor of economics, Syracuse University, Sept. 1, 1975.

John A. Taton: temporary assistant professor, department of economics, University of Georgia, Sept. 1975.

Loren W. Tauer: instructor, department of economics, Iowa State University, Sept. 1, 1975.

Jerry G. Thursby: assistant professor of economics, Syracuse University, Sept. 1, 1975.

Marie C. Thursby, North Carolina State University: assistant professor of economics, Syracuse University, Sept. 1, 1975.

Frank M. Tortora: instructor of economics, Northeastern University, Sept. 29, 1975.

Leonard L. Tsumba: instructor, department of economics, Trinity College, Sept. 1, 1975.

Ronald S. Warren, Jr., University of North Carolina, Chapel Hill: economist, Bureau of Labor Statistics, U.S. Department of Labor, Aug. 1975.

Bernard Wasow, Government Development Bank of Puerto Rico: assistant professor, department of economics, New York University, Sept. 1975.

Daniel H. Weinberg: research associate, National Bureau of Economic Research, Oct. 1, 1975.

William D. White, Harvard University: lecturer/assistant professor of economics, University of Illinois at Chicago Circle, Sept. 1975.

William E. Whitesell, Franklin and Marshall College: regional economist, 3d National Bank Region, Philadelphia.

H. K. Wu, University of Alabama: regional economist, 8th National Bank Region, Memphis.

Diane Zannoni: instructor, department of economics, Trinity College, Sept. 1, 1975.

### *Leaves for Special Appointments*

John S. Akin, University of North Carolina, Chapel Hill: Brookings Policy Fellow, Washington, D.C., July 1, 1975–June 30, 1976.

Mukhtar M. Ali, University of Kentucky: visiting associate professor, Graduate School of Business, University of Chicago, 1975–76.

Michael Balch, University of Iowa: Simon Research Fellowship, University of Manchester, England, 1975–76.

Robert R. Dince, University of Georgia: acting director, department of research and analysis, Comptroller of the Currency, U.S. Treasury, Washington, D.C.

Lehman B. Fletcher, Iowa State University: chief, Agricultural Economics and Sector Planning Division, Bureau of Technical Assistance, USAID, Washington, D.C., to Aug. 31, 1976.

Irving J. Goffman, University of Florida: assistant secretary for income security policy, U.S. Department of Health, Education, and Welfare, Oct. 1, 1975.

Victor P. Goldberg, University of California, Davis: postdoctoral fellow in public choice, Virginia Polytechnic Institute and State University, 1975–76.

Ann R. Horowitz, University of Florida: senior economist, Council on Wage and Price Stability, Washington, D.C., June 1975.

William J. Kelly, University of Georgia: Manpower and Welfare Division, U.S. General Accounting Office, Washington, D.C., Sept. 1975.

C. A. Knox Lovell, University of North Carolina,

Chapel Hill: associate professor, University of British Columbia, Jan 1, 1975–Sept. 1, 1976.

John H. Makin, University of Wisconsin, Milwaukee: visiting associate professor, faculty of commerce and business administration, University of British Columbia, July 1, 1975–June 30, 1976.

Peter B. Meyer, Pennsylvania State University: senior research associate, Correctional Economics Center, American Bar Association, Washington, D.C., winter 1975–76.

Donald O. Parsons, Ohio State University: National Bureau of Economic Research, Palo Alto.

Inderjit Singh, Ohio State University: World Bank, Washington, D.C.

Robert P. Strauss, University of North Carolina, Chapel Hill: staff, Joint Committee on Internal Revenue Taxation, Washington, D.C., Sept. 1, 1975–Aug. 31, 1976.

John C. Weicher, Ohio State University: deputy assistant secretary for economic affairs, U.S. Department of Housing and Urban Development, Washington, D.C.

Daniel Wisecarver, Ohio State University: economic advisor, National Planning Council of El Salvador.

## *Resignations*

Daniel J. Ableman, University of Georgia.

Neil H. Garston, Trinity College: California State University, Los Angeles, Aug. 31, 1975.

Donald J. Harris, University of Wisconsin, Madison: Stanford University.

Richard S. Higgins, University of Georgia: Auburn University, Sept. 1975.

Donald B. Keesing, University of North Carolina, Chapel Hill: International Bank for Reconstruction and Development, June 30, 1975.

John E. Kwoka, University of North Carolina, Chapel Hill: Federal Trade Commission, June 30, 1975.

Martin Landsberg, Trinity College, Aug. 31, 1975.

David G. Tarr, Ohio State University: Bureau of Industry Analysis, Federal Trade Commission, Sept. 1975.

Garrett A. Vaughn, University of Tennessee: Senate Subcommittee on Antitrust and Monopoly.

Obie G. Whichard, University of Georgia: International Investment Division, U.S. Department of Commerce, Sept. 1975.

## NOTE TO DEPARTMENTAL SECRETARIES AND EXECUTIVE OFFICERS

When sending information to the *Review* for inclusion in the Notes Section, please use the following style:

A. Please use the following categories:

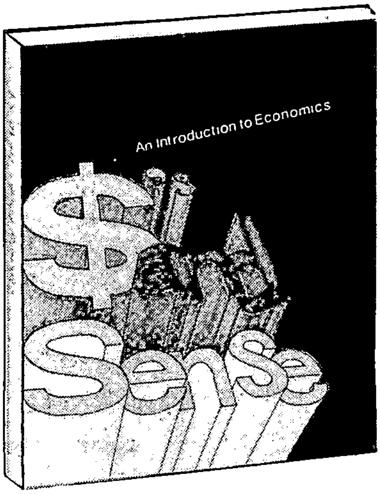
1—Deaths
2—Retirements
3—Foreign Scholars (visiting the USA or Canada)
4—Promotions
5—Administrative Appointments
6—New Appointments
7—Leaves for Special Appointments (NOT Sabbaticals)
8—Resignations
9—Miscellaneous

B. Please give the name of the individual (SMITH, John W.), his present place of employment or enrollment; his new title (if any), his next place of employment (if known or if changed), and the date at which the change will occur.

C. Type each item on a separate 3×5 card, and please do not send public relations releases.

D. The closing dates for each issue are as follows: *March*, November 1; *June*, February 1; *September*, May 1; *December*, August 1.

This announcement supersedes and replaces a letter which was sent annually from the managing editor's office. All items and information should be sent to the Assistant Editor, *American Economic Review*, Box Q, Brown University, Providence, Rhode Island 02912.




# Dollars and Sense
## An Introduction to Economics

Marilu Hurt McCarty
Georgia Institute of Technology

*Dollars and Sense* does *more* with *less.* It covers a balanced selection of macro and micro concepts and applies those principles to many varied real-life situations that are meaningful to students, so that the subject becomes a topic of personal involvement for readers. The result is mastery of the concepts, understanding of how economists apply theory, and practice in independent problem solving.

Concepts and analytical tools are presented simply, followed by examples and applications that relate policies to the individual decision-making level. Each chapter includes learning objectives, self-tests, topics for discussion, and definitions. Special essays on economic issues appear throughout the book. Instructor's Manual with test items.

January 1976, 288 pages, illustrated, paper, $6.95

### National Income Analysis and Forecasting

Edward J. Chambers, University of Alberta
R. Haney Scott, University of Washington
Roger S. Smith, University of Alberta
Complete Instructor's Manual
1975, 406 pages, illustrated, cloth $12.95

### Problems in National Income Analysis and Forecasting
**Revised Edition**

R. Haney Scott, University of Washington
1972, 138 pages, illustrated, paper $4.95

### Macroeconomics
**Analysis and Policy**

James L. Cochrane, University of South Carolina
Samuel Gubins, Haverford College
B. F. Kiker, University of South Carolina
1974, 398 pages, illustrated, cloth $12.95

For further information write to
Jennifer Toms, Department SA
**Scott, Foresman College Division**
1900 East Lake Avenue
Glenview, Illinois 60025

Scott Foresman College Division

# The American Economic Review

## PAPERS AND PROCEEDINGS

OF THE

*Eighty-eighth Annual Meeting*

OF THE

## AMERICAN ECONOMIC ASSOCIATION

*Dallas, Texas, December 28–30, 1975*

MAY 1976

# AMERICAN ECONOMIC ASSOCIATION

Organized at Saratoga, New York, September 9, 1885

## PAST OFFICERS

*Presidents*

Francis A. Walker,* M.I.T., 1886–92
Charles F. Dunbar,* Harvard, 1893
John B. Clark,* Columbia, 1894–95
Henry C. Adams,* Michigan, 1896–97
Arthur T. Hadley,* Yale, 1898–99
Richard T. Ely,* Wisconsin, 1900–01
Edwin R. A. Seligman,* Columbia, 1902–03
Frank W. Taussig,* Harvard, 1904–05
Jeremiah W. Jenks,* Cornell, 1906–07
Simon N. Patten,* Pennsylvania, 1908
Davis R. Dewey,* M.I.T., 1909
Edmund J. James,* Illinois, 1910
Henry W. Farnum,* Yale, 1911
Frank A. Fetter,* Princeton, 1912
David Kinley,* Illinois, 1913
John H. Gray,* Minnesota, 1914
Walter F. Willcox,* Cornell, 1915
Thomas N. Carver,* Harvard, 1916
John R. Commons,* Wisconsin, 1917
Irving Fisher,* Yale, 1918
Henry B. Gardner,* Brown, 1919
Herbert J. Davenport,* Cornell, 1920
Jacob H. Hollander,* Johns Hopkins, 1921
Henry R. Seager,* Columbia, 1922
Carl C. Plehn,* California, 1923
Wesley C. Mitchell,* Columbia, 1924
Allyn A. Young,* Harvard, 1925
Edwin W. Kemmerer,* Princeton, 1926
Thomas S. Adams,* Yale, 1927
Fred M. Taylor,* Michigan, 1928
Edwin F. Gay,* Harvard, 1929
Matthew B. Hammond,* Ohio State, 1930
Ernest L. Bogart,* Illinois, 1931
George E. Barnett,* Johns Hopkins, 1932
William Z. Ripley,* Harvard, 1933
Harry A. Millis,* Chicago, 1934
John M. Clark,* Columbia, 1935
Alvin S. Johnson,* New School, 1936
Oliver M. W. Sprague,* Harvard, 1937
Alvin H. Hansen,* Harvard, 1938
Jacob Viner,* Chicago, 1939
Frederick C. Mills,* Columbia, 1940
Sumner H. Slichter,* Harvard, 1941
Edwin G. Nourse,* Brookings, 1942
Albert B. Wolfe,* Ohio State, 1943
Joseph S. Davis,* Stanford, 1944
I. L. Sharfman,* Michigan, 1945
E. A. Goldenweiser,* Institute for Advanced Study, 1946
Paul H. Douglas, Chicago, 1947
Joseph A. Schumpeter,* Harvard, 1948
Howard S. Ellis, California, 1949
Frank H. Knight,* Chicago, 1950
John H. Williams, Harvard, 1951
Harold A. Innis,* Toronto, 1952
Calvin B. Hoover, Duke, 1953
Simon Kuznets, Pennsylvania, 1954
John D. Black,* Harvard, 1955
Edwin E. Witte,* Wisconsin, 1956
Morris A. Copeland, Cornell, 1957
George W. Stocking,* Vanderbilt, 1958
Arthur F. Burns, Columbia, 1959
Theodore W. Schultz, Chicago, 1960
Paul A. Samuelson, M.I.T., 1961
Edward S. Mason, Harvard, 1962
Gottfried Haberler, Harvard, 1963
George J. Stigler, Chicago, 1964
Joseph J. Spengler, Duke, 1965
Fritz Machlup, Princeton, 1966
Milton Friedman, Chicago, 1967
Kenneth E. Boulding, Colorado, 1968
William J. Fellner, Yale, 1969
Wassily Leontief, Harvard, 1970
James Tobin, Yale, 1971
John Kenneth Galbraith, Harvard, 1972
Kenneth J. Arrow, Harvard, 1973
Walter W. Heller, Minnesota, 1974
Robert Aaron Gordon, California, 1975

*Secretaries*

Richard T. Ely,* 1886–92
Edward A. Ross,* 1893
Jeremiah W. Jenks,* 1894–96
Walter F. Willcox,* 1897–99

*Treasurers*

Edwin R. A. Seligman,* 1886–90
Frederick B. Hawley,* 1891–95
Charles H. Hull,* 1896–99

*Secretary-Treasurers*

Charles H. Hull,* 1900
Frank A. Fetter,* 1901–06
Winthrop M. Daniels,* 1907–08
Thomas N. Carver,* 1909–13
Allyn A. Young,* 1914–20
Ray B. Westerfield,* 1921–25
Frederick S. Deibler,* 1925–35
James Washington Bell,* 1936–61
Harold F. Williamson, 1962–70

*Editors*

Davis R. Dewey,* 1911–40
Paul T. Homan,* 1941–51
Bernard F. Haley, 1952–62
John G. Gurley, 1962–68

*Deceased

## OFFICERS FOR 1976

*President*

Franco Modigliani
Massachusetts Institute of Technology

*President-elect*

Lawrence R. Klein
University of Pennsylvania

*Vice-Presidents*

Barbara R. Bergmann, University of Maryland
Harry G. Johnson, University of Chicago

*Secretary*

C. Elton Hinshaw, Vanderbilt University

*Treasurer and Editor of the Proceedings*

Rendigs Fels, Vanderbilt University

*Managing Editor of the American Economic Review*

George H. Borts, Brown University

*Managing Editor of the Journal of Economic Literature*

Mark Perlman, University of Pittsburgh

*Elected Members of the Executive Committee*

Andrew F. Brimmer, Harvard University
Paul M. Sweezy, Monthly Review
Carolyn Shaw Bell, Wellesley College
Burton A. Weisbrod, University of Wisconsin
Edmund S. Phelps, Columbia University
Alice M. Rivlin, Congressional Budget Office

*Ex Officio Members*

Walter W. Heller, University of Minnesota
Robert Aaron Gordon, University of California, Berkeley

*Counsel*

Donald F. Turner, Harvard University

# THE AMERICAN ECONOMIC REVIEW

VOL. 66 | MAY, 1976 | NUMBER 2

*PAPERS AND PROCEEDINGS*

OF THE

*Eighty-eighth Annual Meeting*

OF THE

AMERICAN ECONOMIC ASSOCIATION

*Dallas, Texas*

*December 28–30, 1975*

*Program Arranged by* FRANCO MODIGLIANI

*Papers and Proceedings Edited by* RENDIGS FELS AND C. ELTON HINSHAW

PRINTED BY GEORGE BANTA COMPANY, INC.

*Publication Office:* Menasha, Wisconsin 54952

*Executive Office:* American Economic Association
1313 21st Avenue South, Nashville, Tennessee 37212

Correspondence relating to the *Papers and Proceedings,* the *Directory,* advertising, permission to quote, business matters, subscriptions, membership and changes of address should be sent to the Secretary, C. Elton Hinshaw, 1313 21st Avenue South, Nashville, Tennessee 37212. Change of address must reach the Secretary at least six (6) weeks prior to the month of publication. The Association's publications are mailed second class.

THE AMERICAN ECONOMIC REVIEW, including four quarterly numbers, the *Proceedings* of the annual meetings, the *Directory,* and Supplements, is published by the American Economic Association and is sent to all members five times a year, in March, May, June, September, and December. Association membership dues for 1976, which include a subscription to both The *American Economic Review* and the *Journal of Economic Literature,* are as follows: $25.00 for regular members with rank of assistant professor or lower, or with annual income of $12,000 or less; $30.00 for regular members with rank of associate professor, or with annual income of $12,000 to $20,000; $35.00 for regular members with rank of full professor, or with annual income above $20,000; $12.50 for junior members (registered students). Certification must be submitted yearly. Subscriptions (libraries, institutions, or firms) are $37.50 a year. Only subscriptions to both publications will be accepted. Single copies of either journal may be purchased from the Secretary's office, Nashville, Tennessee. For foreign postage, add: PUAS countries, $3.20; all other countries and Canada, $3.75.

Second-class postage paid at Nashville, Tennessee and at additional mailing offices. Printed in U.S.A.



Price $10.50

# CONTENTS

THE purpose of the American Economic Association, according to its charter, is the encouragement of economic research, the issue of publications on economic subjects, and the encouragement of perfect freedom of economic discussion. The Association as such takes no partisan attitude, nor does it commit its members to any position on practical economic questions. It is the organ of no party, sect, or institution. Persons of all shades of economic opinion are found among its members, and widely different issues are given a hearing in its annual meetings and through its publications. The Association, therefore, assumes no responsibility for the opinions expressed by those who participate in its meetings. Moreover, the papers presented are the personal opinions of the authors and do not commit the organizations or institutions with which they are associated.

# RICHARD T. ELY LECTURE

# Stabilization Policy in Open Economies with Endogenous Politicians

*By* ASSAR LINDBECK*

Stabilization policy has since Keynes been based on two main premises: that macroeconomic fluctuations are largely a result of instability in market behavior; *and* that national aggregate demand management, possibly combined with some restraints in wage bargaining, is the basic means of overcoming these fluctuations.

In this lecture, the emphasis will be somewhat different both in describing the mechanisms of macroeconomic fluctuations and in designing policies to reduce them. Specifically, macroeconomic fluctuations will be seen in the context of a complicated *interaction* between market forces and government behavior, i.e., as an interplay between the economic and the political systems. Moreover, when discussing market forces, the emphasis will be on international rather than national forces.

With respect to policy prescriptions, I will argue that aggregate demand management, though important, is insufficient for a successful stabilization policy. Particularly in distinctly open economies with rather high levels of capacity utilization, there are at least four additional, potentially important types of national stabilization policy: (1) policies designed to influence relative prices; (2) market-improving and mobility-increasing policies; (3) supply management; and (4) selective demand management.

Several of these policies would, if implemented, result in some redesign of the structure of the economic system. Beyond this, however, as government actions often contribute to macroeconomic fluctuations, rather than dampening them, some redesign of the national and international political system too is crucial for an improvement of macroeconomic stability.

The exposition will be organized around three basic, highly interrelated problems of national stabilization policy: (I) the complications related to the international influences on the domestic economy—i.e. "the open economy problem"; (II) the limitations of aggregate demand management; and (III) the tensions between, on the one hand, the political process and, on the other hand, the requirements of a successful stabilization policy.

I shall begin with a technical economic analysis of the first two issues—the open economy problem and the limitations of aggregate demand management. In the last part of the lecture, the problems of the economic system will be explicitly related to the functioning of the political system, and hence stabilization policy be analyzed in the context of the interrelationship between the two systems.

## I. The Open Economy Problem

It is useful to approach the problem about international influences by way of a

* Professor of International Economics and Director of the Institute for International Economic Studies, University of Stockholm. I am indebted to Franco Modigliani and to my colleagues at the Institute for valuable comments to earlier drafts.

model of an extremely open economy and thereafter to investigate how the conditions for stabilization policy will change when the degree of openness is gradually reduced. In particular, I want to show how the relative roles of, on the one hand, *aggregate demand management* and, on the other hand, policies designed to influence *relative prices* depend on the type and degree of openness of the national economy. At the same time, the analysis brings out how the relative efficacy of monetary and fiscal policy depends on the openness of the economy.

### A. An Ultra-Open Economy

Thus, let us start with a mythological "ultra-open" economy model, where (1) all "goods"—commodities, labor and financial assets—are *tradables*, and (2) domestic and foreign goods are *perfect substitutes*. Assuming, to begin with, that exchange rates are fixed and that all goods move freely, domestic prices (excluding indirect taxes and subsidies) of commodities, labor, and financial assets are determined by supply and demand on world markets rather than on national markets. Let us further assume that the country and its firms are so small relative to the outside world that world market prices for all goods are independent of the economic actions of the country itself—the celebrated "small country assumption."

It follows immediately from these assumptions that the room for national stabilization policy is extremely limited in such an economy. Domestic unemployment policy is clearly impossible for the national government, as the labor market is international rather than national. Thus, aggregate demand management is pointless. However, the commodity price level can be influenced by the exchange rate, as well as by tariffs and indirect taxes or subsidies on foreign trade. That is basically all that national stabilization policy can do in this model.[1]

### B. Four Steps Towards the Real World

This caricature of open economies is perhaps a useful approximation for geographical regions within some countries. Let us, however, move somewhat closer to the national economies of the real world. Four steps will be taken in this direction, each one reducing the degree of "openness" of the domestic economy.

*The first step* is to drop the assumption that the labor supplies in various countries are perfect substitutes for each other and perfectly mobile; the labor supply curve that confronts the domestic production sector is then upward sloping rather than horizontal. We may now talk about a *domestic* labor market, though connected with foreign markets. The main implication for economic policy, as compared to the previous model, is that variations in the exchange rate, tariffs, and indirect taxes or subsidies on foreign trade and production may now be used to influence the domestic unemployment level and not just the price level. However, a condition is that the money wage rate does not adjust completely to these policy changes, within the period under consideration, so as to restore the initial real wage rate that confronts firms. Some stickiness of wages in the model is both a reason for unemployment and a prerequisite for influencing output and unemployment by way of exchange rate policy and indirect taxes or subsidies on trade or domestic production. The government can, of course, also reduce unemployment and increase domestic output by

[1] Whereas *unemployment* cannot be influenced, the level of employment and output in the domestic production sector may be affected by way of indirect taxes (or tariffs) and subsidies on foreign trade and domestic production.

employing labor directly and producing output itself.

However, *aggregate demand management* in the commodity and asset markets is still useless. All variations in domestic commodity demand, at a given price level fully "spill over" into the trade account of the balance of payments without any effects on prices or on domestic production and employment. Similarly, all variations in asset demand by the authorities, i.e., open market operations, fully spill over into the capital account of the balance of payments without any effects on interest rates or on domestic private holdings of financial assets and money.

To summarize, in this second model stabilization policy has to rely entirely, except for direct government employment of labor, on policy measures that change *relative prices,* more specifically the real wage rate that confronts firms. In other words, while in this model monetary policy (except exchange rate changes) is useless for domestic policy purposes, there are several types of fiscal policy measures that are potentially useful for influencing the domestic unemployment level (and output): indirect taxes or subsidies on trade and production, as well as direct government demand for labor.

*The second step* toward the real world is to drop the assumption that domestic and foreign tradable commodities are perfect substitutes, which means that total demand (domestic plus foreign) for domestic output can now be represented by a downward sloping aggregate demand curve rather than by a horizontal price line. (The implications are rather similar to dropping the small country assumption for the commodity market.) As a consequence, domestic aggregate demand management, by way of conventional fiscal policy, now becomes a useful tool for influencing domestic output and unemployment—in addition to policy measures that change relative prices. Moreover, if not only wages but also prices are sticky downward, as they are in the real world, ordinary Keynesian-type "demand failures" with unemployment disequilibrium may take place, which increases the scope for aggregate demand management even further.

If, as *a third step,* nontradable commodities are introduced in the model, the national autonomy of demand management by way of fiscal policy is further increased—for domestic unemployment as well as the general price level—since there is no foreign component at all in the demand function in this market.

Monetary policy is, of course, still useless as a tool of demand management for domestic purposes, so long as we stick to the assumption that all financial assets are tradables and that foreign and domestic assets are perfect substitutes (at fixed exchange rates). If this doubtful assumption (which seems to have been taken as a realistic description of the real model by many theoretical model-builders) is removed, as *a fourth step* towards the real world, monetary policy too will be useful as a tool of demand management. More generally, open market operations will be more effective in influencing the domestic economy the less the substitutability between domestic and foreign financial assets is, as compared to the substitutability between domestic financial assets and money. The reason is that the lower the former substitutability, the smaller the fraction of open market operations that will "leak" abroad, thereby offsetting the effects on domestic money and credit markets of open market operations. In the real world, both specific institutional conditions in the credit market and the use of nonconventional monetary and fiscal policy instruments, mainly tools that create wedges

between the rate of return of foreign and domestic assets, can significantly increase the national autonomy of stabilization policy even in otherwise rather open economies.

*C. The Relative Openness of Markets*

What has been said so far illustrates the fact that the less open the economy and the less flexible domestic wages and prices, the more useful aggregate demand managment will be as a complement to policy measures that change relative prices (such as exchange rates and indirect taxes or subsidies). The discussion has also underlined that the *relative* openness of different markets—for labor, commodities and credit—is crucial for the efficacy of various policy tools. The relative openness of the labor and commodity markets has some additional implications with respect to the performance of stabilization policy. For instance, the rather realistic assumption that the commodity market is more open than the labor market yields some important conclusions.

First, international inflation would then be expected to be transmitted among nations *largely* by way of a "direct price influence" on tradables and possibly also an indirect price influence on nontradables via wage formation. The more important this direct and indirect "price link" is, the less important the multiplier effects and the monetary effects via the balance of payments should be. This conclusion rests, of course, on the assumption that prices for domestic tradables are primarily determined in the world market rather than by domestic economic activity and hence domestic demand and supply conditions. The inference, in fact, is consistent with the observed tendency in recent years for inflation also to be transmitted internationally to countries with considerable domestic slack (due, for instance, to restrictive domestic demand management). It is also consistent with the experience in macroeconometric model building that it has been difficult in many countries to identify the excess demand situation in the domestic commodity market as an important factor behind the rate of domestic price change. Thus, the apparent limited sensitivity of domestic prices to *domestic* demand conditions may not be a result only of rigidly "administered prices"; the geographical market domain may partly be misspecified in the models.

Secondly, the real wage rate and therefore also the rate of return on physical assets in the tradable sector, as well as the relative price between tradables and nontradables, can easily become "out of line" in economies where the commodity market is highly open relative to the labor market. The reason is that whereas the price level for tradables is determined primarily in the world market, nominal wage rates are directly determined by domestic circumstances, such as by excess demand in the national labor market, by domestic price expectations, and by specific national circumstances for wage bargaining, which are often in fact related to the national political situation.

Thus, an "improper" real wage rate, with a concurrent low profitability level, can be expected to be a more frequent reason for unemployment, low investment incentives, and a current account deficit in economies where the commodity market is open relative to the labor market than in economies where the degree of openness is about the same in both markets. This means that policies designed to influence the real wage rate will be more important for investment, employment, and the current account, the more open the commodity market is relative to the labor market. As an empirical generalization, I think it is safe to say that stabilization policy has suffered considerably in many countries from a neglect of relative prices

—such as the real wage rate, the relative price between tradables and nontradables, and in particular profitability—as compared to aggregate demand. The United Kingdom is an obvious example.

This means that whereas for a closed economy *any* point on the long-run Phillips curve may be regarded as an equilibrium point, there is in principle only *one* equilibrium point on the long-run Phillips curve for an open economy in the sense that only one point is consistent with both constant inflation and an unchanged position for the current account of the balance of payments (normally at the point where domestic and foreign cost increases are equal). With floating exchange rates, any point on the long-run Phillips curve can in principle be chosen, just as in a closed economy.

A third inference from this analysis is that it does not make much sense for small countries to fight inflation by cutting back domestic aggregate demand (rather than revaluing) if inflation is brought about by international price changes. In this specific sense, the government in a small country can often tie aggregate demand management to unemployment and current accounts targets, rather than to the price target. This was, in fact, what most small countries did in the 1974-75 stagflation period, when they let the large countries alone fight world inflation by way of demand-reducing polices and drastically lower capacity utilization. Most small countries could, and did, take "free rides" on the anti-inflationary policies of the big countries—as long as the governments in the small countries were willing to accept a deterioration in the current account of the balance of payments. Perhaps it should be added that having taken advantage of the free ride, the small countries could then boast how much more successful they were than the large countries in sustaining high employment!

Let us also combine the earlier assumption about a basic difference in price formation behavior between commodities and labor, and between tradables and nontradables, with an assumption about a difference in the rate of productivity increase between these two commodity sectors. Let us also add an *accommodating economic policy on the demand side*. It is then possible to derive rather strong theorems about antiinflationary policies in an individual country.

*First of all*, recommendations in incomes policy discussions have usually been based on the assumption that price stability can be achieved by restricting the rate of nominal wage increase for the economy as a whole to the average rate of productivity increase for the economy. This recommendation has to be strongly modified in an open two-sector model of the type discussed here if the rate of productivity increase is smaller in the nontradable than in the tradable sector, and if wages tend to develop in a parallel fashion in the two sectors. For prices of nontradables would then rise in spite of a proportional development of wage rates and productivity for the economy as a whole, and there is no reason why prices would *fall* in the tradable sector, to "compensate" for the price rise for nontradables, in a fixed exchange rate regime. The conclusion is, of course, that continuous revaluations are necessary for stability of the general price level, even if world market prices for tradables are constant and the rate of wage increase in both sectors is equal to the increase in productivity for the economy as a whole.

*Secondly*, another widely heard recommendation for fighting inflation has been to step up the rate of productivity growth, particularly in manufacturing, for instance by way of harder work, higher investment, and more rapid improvements in technology. However, a faster rate of productivity increase for tradables would, in an

open two-sector model of the type discussed here, tend to *increase* the rate of inflation as long as the exchange rate is fixed. The reason is that the "room" for wage increase (at constant profit margins) in the tradable sector would then go up, generating a more rapid wage increase. When these wage increases are spread to the nontradable sector, the rate of price increase there will go up without any "corresponding" fall in the rate of price increase in the tradable sector, where prices are tied to the world market. By contrast, a faster rate of productivity increase for nontradables, such as for many services, would clearly help reduce inflation.

In fact, in the extreme case where (1) prices for tradables are "parametrically" determined by the outside world, (2) profit margins are constant, both in the tradable and the nontradable sector, and (3) a "cost-accommodating" policy is pursued on the demand side, the average price level would increase by the *sum* of the rate of the exogenous price increase for tradables and the difference in the rate of productivity increase between the two sectors, the latter variables being weighted by the share of nontradables in the overall price index.[2] A model of this kind corresponds to what in the international literature has been called "The Scandinavian model of inflation" (Aukrust; Edgren-Faxén-Odhner). Of course, countries tending to inflate more rapidly than this can do so by devaluing from time to time, and countries tending to inflate less can do so by concurrent revaluations.

### D. Flexible Exchange Rates

The last points illustrate the usefulness in macroanalysis of a disaggregation between tradables and nontradables and of looking at the relative price between these. The analysis also illustrates the well-known fact that exchange rate flexibility, either through discretionary changes or freely floating rates, is a *necessary* condition for a reasonably independent national price trend. It is also clear that exchange rate flexibility can partly shield the economy from volume fluctuations on world markets by appreciations in world booms and depreciations in world recessions—occasionally an example of a "beggar-thy-neighbor policy."

However, the degree to which this shielding from international price and volume fluctuations takes place depends both on the type of international disturbance and on the monetary and fiscal policy pursued in the country in question. For instance, the economy may be shielded rather well from proportional increases in world prices for *all* tradables by way of an appreciation of the exchange rate in proportion to the international price increase. On the other hand, it is well known that different rates of price change for different products (including terms-of-trade changes) are less effectively handled by exchange rate changes, since stability of the general price level then requires

[2] Make the following notations:

$p_w$ = rate of price change for domestic tradables on world market;

$p_T$ and $p_N$ = ditto for tradables and nontradables, respectively, on domestic market;

$e$ = rate of change of exchange rate;

$w_T$ and $w_N$ = rate of wage change in tradable and nontradable sector, respectively;

$q_T$ and $q_N$ = rate of productivity increase in tradable and nontradable sector, respectively;

$\alpha$ = weight in price index for tradables.

The basic *assumptions* of the model then are:

$$p_T = p_w + e; \qquad w_T = p_w + e + q_T = w_N;$$
$$p_N = w_N - q_N = p_w + e + q_T - q_N.$$

The *inference* for the rate of change of the price level is:

$$\begin{aligned} p &= \alpha \cdot p_T + (1-\alpha)p_N \\ &= \alpha(p_w + e) + (1-\alpha)(p_w + e + q_T - q_N) \\ &= (p_w + e) + (1-\alpha)(q_T - q_N). \end{aligned}$$

Thus, the domestic price index will rise by the *full* value of $p_w$ and $e$, and by the weighted difference between $q_T$ and $q_N$.

reductions in prices for some commodities to compensate for price increases for others, and this may be difficult to achieve in a world with limited price flexibility downwards.

In addition, in a *floating* rate system, isolated import price increases may easily lead to market-induced depreciations, which tend to accentuate the inflationary price influence from abroad—as in Japan and the United Kingdom after the oil price increase. Moreover, both theoretical considerations and empirical evidence suggest that it is *portfolio shifts* that, in the short run, dominate exchange rate developments in a floating exchange rate regime (and reserve movements in fixed rate systems). (See, for instance, P. J. K. Kouri.) Thus in a regime of floating rates, shifts in portfolios by foreign and domestic asset holders may create considerable disturbances of the domestic price level, as well as of profitability and the allocation of resources, via induced changes in exchange rates, particularly if price increases are (partly or wholly) irreversible.

Basically, much of this simply illustrates the Jan Timbergen rule of economic policy, according to which two targets cannot as a rule be achieved with one policy instrument. In general, we should not expect a given exchange rate change to bring about the desired development of *both* the balance of payments *and* the price trend, if other policy instruments are not adjusted accordingly. More specifically, if money wages react strongly and rapidly to exchange rate changes brought about either by official interventions or by portfolio shifts, it is the domestic price level, rather than the balance of payments, that is affected by the change in the exchange rate. If, by contrast, money wages do *not* respond much to exchange rate changes, these will influence the balance of payments rather than the price level. In general, the more efficient exchange rate changes are in influencing the balance of payments, the less efficient they are in influencing the price trend, and vice versa. Thus, what will actually happen in a specific situation to the price level and the balance of payments after a change in the exchange rate depends both on the system of wage formation and on the demand management policy that is actually pursued, including the monetary policy.

## II. Limitation of Aggregate Demand Management

A main conclusion of the previous analysis was that a successful stabilization policy *in open economies* cannot rely only on aggregate demand management; relative prices between tradables, nontradables, and labor are also highly relevant. The case for a heavy reliance on policy measures that change relative prices can be put even stronger if three additional limitations of aggregate demand management are emphasized: (1) the importance of substitutions between periods for stabilization policy; (2) problems connected with the heterogeneity of the commodity and labor markets; and (3) difficulties related to competing claims of various organized groups in society. Let us look at these three aspects in turn—still neglecting the complications for stabilization policy that are embedded in the political system.

### *A. Substitutions between Periods*

Since a main purpose of stabilization policy is to change the *timing* of spending, policy tools that create substitution effects between periods via changes in relative prices between periods are particularly useful. It is from this point of view that interest rate changes have always been regarded as appropriate for stabilization policy, the interest rate being a relative price between variables in different periods. However, similar ef-

fects can be obtained by *temporary* changes in indirect taxes or subsidies on spending and/or production—of consumption as well as investment. Some European countries have, in fact, experimented with such measures—with some success.[3] Indirect taxes and subsidies of this type also have the advantage of "striking directly" on what we want to affect—the *flow* of spending on investment or consumption—rather than, like interest rate changes, on the market value of the *stock* of earlier issued financial assets. Furthermore, indirect taxes and subsidies can be used also when the external constraint limits the use of interest rate policy.

While changes in *direct* taxes on households that are expected to be temporary may have very small effects on spending, due to the fact that consumption in one period is a function of the path of past and expected future income, *temporary* indirect taxes can overcome this deficiency by creating substitution effects between periods (Lindbeck 1963, Chaps. IV and V). To achieve similarly strong short-term effects by way of temporary changes in direct taxes or transfers, very large doses may be necessary.

If stabilization policy in major countries during the 1974-75 recession had relied more on temporary reductions of indirect taxes or temporary increases in subsidies on consumption as well as investment, stabilization policy would probably have been much more successful in stimulating the economies without accelerating inflation—assuming that this really was what the governments wanted. In fact, reductions in indirect commodity taxes or increases in indirect subsidies seem ideally suited in situations of "stagflation," as such measures both limit inflation (of market prices) and reduce unemployment.

[3] Norway, Sweden and West Germany are examples. For an account of the Swedish experience, see Lindbeck 1974.

### *B. Heterogeneity of Commodity and Labor Markets*

Another obvious weakness of "pure" *aggregate* demand management is that excess demand in the commodity and labor markets is often very unevenly distributed between various submarkets. A policy that tries to realize full employment by way of aggregate demand management only is therefore likely to run into bottlenecks and higher inflation long before "full" capacity utilization is reached for the weakest sectors of the economy.

If we are not satisfied either with a rather low average level of capacity utilization or with the rate of inflation, in fact possibly an accelerating one, that follows from an economy with excess demand in many submarkets, there seem to be at least three alternative strategies—still neglecting the mutual interaction between the economic and the *political* systems.

*One strategy* is increased factor mobility of labor as well as capital. In the case of labor, this would mean greater flexibility of relative wages and cheaper and better organized systems for labor exchange and retraining and other labor-mobility-increasing features. I will not dwell on this well-known and rather generally accepted point.

*A second strategy* would be "supply management" as a complement to demand management. The basic idea would be to stimulate capacity increases in sectors that are likely to be bottlenecks in the next boom. This point is relevant at least for some nontradables, perhaps mainly for the public infrastructure and sectors where the government controls prices and/or output. In many cases, simply a removal of government regulations, which often distort relative prices and create bottlenecks, may be the most efficient type of

supply management. A slightly more interventionist version of supply management would be to try to "flatten out" the short-run supply curves for goods and services by subsidies for the employment of high-cost labor (such as employees with poor training or with various handicaps)—and perhaps also to give temporary subsidies to the new hiring of labor in general, in particular during a recession. The idea is, of course, to limit the size of the increase in market prices—via a devaluation or increased domestic aggregate demand—that is necessary to induce increased output and employment. (See Rehn.)

*A third strategy* would be "selective demand management" in order to adjust the composition of demand to the composition of productive capacity between sectors and regions. The idea is, again, to avoid bottlenecks, largely by a conscious manipulation of relative prices, when higher capacity utilization is approached.

In general, we could perhaps say that in the same way as rigidities of the average wage and price *levels* create a need for aggregate demand management, rigidities in *relative* wages and prices make a case for mobility-promoting policies, and selective supply and demand management. As criteria for the "composition" of such short-run selective demand and supply management could perhaps be used the distribution between sectors and regions of vacancies, unused capacity, inventories, unfilled orders, delivery times, etc.

A more specific argument for such policies is that macroeconomic disturbances may occur at different points in the macroeconomic system and that it may be better to counteract these disturbances where they take place, rather than to try by way of broadly based demand policies to prevent them from influencing macroeconomic target variables (Lindbeck, 1963, pp. 151-73).

Policies of these types have to be classified under the nowadays rather dirty word "fine tuning." However, it is important to stress that there are at least two different kinds of "fine tuning."

*One* type is that the government tries continuously to keep aggregate demand very close to the capacity ceiling of the economy, which tends to create a situation with excess demand ("bottlenecks") in many sectors simultaneously with excess supply ("slack") in others. It was this type of fine tuning that ran into severe problems during the late 1930's and early 1970's.

A *different* type of fine tuning, which has been discussed here, is to be very "modest" with expansions of aggregate demand after a rather high, but not quite "full" level of capacity utilization has already been reached, and instead try to achieve full employment by way of factor mobility policies and selective demand and supply management. This type of "selective fine tuning" is, in fact, quite consistent with the theory of "the natural rate of unemployment," as the policy implication of that theory is that attempts to remove unemployment, below a certain level, by way of increased aggregate demand, is self-defeating. It is, of course, an open question, in practice, whether full employment policies by way of selective fine tuning could have avoided the acceleration of inflation in the world that followed the continuously higher average level of capacity utilization in the Western world from the mid-1950's to the late 1960's and early 1970's.

### C. Competing Claims

Another macroeconomic aspect of relative prices is that *organized groups* in our present day economies—labor unions, farmers' organizations, employers' organizations, large corporations, and public authorities—try to use their market power to change relative prices in their

favor in the struggle for income shares and relative factor returns. It is well known that aggregate demand management may not eliminate inflation due to such competing claims, unless the policy is extremely restrictive and unemployment-creating.

In cases like these, inflation should not be regarded as an "effect" of a change in some *specific* variables—such as "exogenous" spending, open market operations, or the quantity of money—but rather as a common *process of adjustment* of the economy to various types of inconsistent claims, with some groups being disappointed at every point of time. The process presupposes, of course, a "cost-accommodating" demand management policy including an accommodating supply of money and credit.

A specific and increasingly important aspect of the struggle for income shares is the inconsistency of claims by the public and private sector. For instance, it has been noted in many countries that labor unions often ask for "compensation"—in the form of wage increases—to offset tax increases designed to finance increased public spending. This means that attempts by the authorities to fight demand inflation are likely to result in cost-push inflation instead. What makes this conflict ironic is that the labor unions which try to compensate themselves for tax increases are often the very same pressure groups that lobby for the increased government spending that has made the tax increases necessary.

If this type of compensation mechanism operates also for *automatic* tax increases, i.e., when tax revenues increase at given tax rates due to higher national income, the theory of automatic stabilizers on the demand side has to be reformulated to allow for the destabilizing effects on the cost side. When wage demands are based on such ambitions to protect after-tax real income, quite explosive wage-price spirals may be created in countries with highly progressive tax systems—if aggregate demand management is "permissive" in order to prevent unemployment. The basic reason for these problems seems to be the same as why deteriorations in the terms of trade may create domestic wage-price spirals of "cost-push type": real income resistance of income earners. Thus, conflicts about *relative* incomes and wages are not the only reason for spirals of this type; concern and disappointment about *real* disposable incomes may also generate such spirals.

If the progressiveness of the tax system is very high, it is easy to show that the only chance for employees to obtain increases in their real disposable wages may be to accept *reductions* in nominal wages and hence "cheat" the government on real disposable income by "exploiting" the progressiveness of the tax system "in the reverse direction," and as a consequence lowering the public share of national income (Calmfors-Lundberg, pp. 122-29; Lindbeck 1974). However, this is certainly very difficult to bring about in practice, as it presupposes both collusion between most groups of income-earners, and downward price flexibility.

The main conclusion of these considerations is probably that *both* widespread dissatisfaction concerning the distribution of income *and* a rapid expansion of the public sector, in particular in the context of an extremely progressive tax system, in reality are difficult to reconcile with reasonably stable prices. This is hardly a comfortable conclusion.

## III. The Political System: Endogenous Politicians

The previous analysis has been somewhat "non-Keynesian" in its emphasis on relative prices and profitability and not just (or mainly) on aggregate demand. However, the analysis has been in line with conventional Keynesian philosophy

in the sense that it has been based on the view that we live in an unstable market system, which could and should be stabilized by conscious economic policies of national governments. Such recommendations, however, neglect instabilities and imperfections in the political-administrative system. This is clearly a nonsymmetrical treatment of conceivable deficiencies in the market system and in the political-administrative system.

This does not mean that we have to swing to the other extreme, and assume, like the Chicago school, that the world is just the other way around: that we live in a highly stable market system which from time to time is upset by destabilizing actions taken by governments. Instead, I think that the most reasonable approach, at the present state of knowledge, is to assume that there are instabilities and imperfections in *both* the market system *and* the political-administrative system, and that these interact in a complex way. It is obvious that macroeconomic fluctuations today are so intimately connected with government policies that realistic explanations and forecasts of macroeconomic fluctuations require that government behavior be analyzed as an integral part of the fluctuations. For instance, it seems that the differences between various short-term economic forecasts today depend less on divergent views on the functioning of the private sector than on different assumptions about the future economic policy. An important aspect of such assumptions is various hypotheses as to how governments most likely will react to future economic events. This means that it is useful to treat the government as an *endogenous* rather than an exogenous variable in the macroeconomic system.

### *A. Political Behavior Functions*

Granting the need to "endogenize" politicians, it is obvious that this can be achieved in several alternative ways. One approach to the positive theory of political behavior is to assume that politicians are well informed and idealistic "guardians of the general good" (as defined, perhaps, by the ideology of the government). This would mean that the *normative* behavior functions derived from the welfare theory of economic policy—as developed by Tinbergen, James E. Meade and others—would be interpreted also as *positive* behavior functions of politicians.

An alternative approach to the theory of political behavior is to take the opposite position and assume, like Anthony Downs, that politicians are concerned with their own welfare, obtained by seizing and staying in power, rather than with the welfare of society as a whole. More specifically, a politician in a democracy is then, to quote Downs, regarded as "an entrepreneur selling policies for votes instead of products for money" (Downs, p. 137). Analytically, behavior functions for politicians could then be derived by replacing an "idealistic" preference function of politicians à la Tinbergen-Meade with a vote-getting function, which is maximized subject to the constraints defined by the trade-off possibilities between the various vote-creating variables.

A more "eclectic," and I think more realistic, approach would be to combine the idealistic and the popularity approaches, with the relative weights between the idealistic variables and the popularity variable changing systematically over time during the course of an election period—the popularity variable gaining weight relative to the idealistic variables when an election day is approached, which would mean that the Downs theory would be approached when the election comes closer. In an explicitly intertemporal model we could, with some exaggeration, say that as the time to the election approaches zero, the discount rate for the future economic variables goes to infinity. Every government wants, of course, to be-

lieve that this type of behavior coincides with what is "the general good for society as a whole" *in the long run*.

On the basis of this eclectic approach to the theory of economic policy behavior, three sets of relations have to be specified: (1) the properties of the asserted target-preference function of the government, i.e., a specification of both the idealistic and the popularity variables in the function, and their relative weights at different points of time; (2) a specification of how the popularity variable is influenced; and finally (3) a specification of the economic constraints within which the maximization of the target-preference function is supposed to take place, in particular a description of how the government can influence the economy. (For approaches along these lines, see for instance Frey-Lau; Lindbeck 1973).

*B. Endogenous Stabilization Policy*

When *stabilization policy* is discussed, it is probably important to include in the target-preference function at least four "idealistic" variables, each one characterized by falling marginal utility, or rising marginal disutility: real disposable income of households, unemployment, inflation, and the current account of the balance of payments (or the stock of reserves)—and most likely also the rate of change of these variables. When choosing an appropriate popularity variable in the preference function, several alternatives are possible. An expression for party sympathies among the electorate is perhaps a reasonable variable, operationally defined for instance as "sympathy" answers to Gallup polls.

It is likely that the popularity variable is influenced by approximately the same variables as we have chosen to treat as idealistic variables in the target-preference function, though possibly with different weights and different timing. In particular, I will make two assumptions which make a preference function that includes both idealistic variables and a popularity variable differ substantially from one which includes only the idealistic variables: (1) *Recent* events are assumed to be considerably more important for the popularity of the government than events long ago—an assumption about a "short memory" of voters. (2) The electorate is assumed to have very imperfect knowledge about the *future* effects of policy actions, such as effects that emerge after the next election.

That voters are interested in their real income and possibly its rate of change is, of course, consistent with conventional micro theory of household behavior. Moreover, a number of empirical studies indicate that governments tend to gain popularity in periods of prosperity and rising real income and lose it in periods of depression and falling (or only slowly rising) real income (Kramer; Åkerman).[4] Considering the important role, in the political discussion after the Second World War, of inflation, unemployment, and the current account of the balance of payments (or possibly the level of exchange reserves), and their rates of change, we should also expect these variables to be important for the popularity of a government. In fact, several empirical studies suggest that this is the case (Lepper; Nordhaus).

On the basis of these assumptions, we would expect two characteristic features of government behavior over the cycle.

*First*, due to the assumption of increasing marginal disutility of inflation, unemployment, and current account deficits, we would expect shifting emphasis on the various macroeconomic policy targets as

[4] There is an identification problem here, however, in countries where the government can choose election time: elections may be held in periods of rising real disposable income, rather than the other way around. Or both may have been deliberately influenced by the government.

these variables happen to change themselves due to exogenous disturbances—even without consideration of the popularity variable and also without having to assume any shift in the target-preference function. Inflation, current account (or balance-of-payments) deficits, and unemployment will succeed each other as "the Enemy Number One of the Country," thus generating a stop-go policy cycle in response to factual events.

*Secondly,* as soon as we also take account of the popularity variable in the target-preference function, we would expect that the closer we are to an election, the more inclined the government is to take expansionary actions. This conclusion follows already from the introduction of real disposable income, and/or its rate of change, in the popularity function, because of the ease and speed by which governments can "buy" popularity, and hence votes, simply by raising temporarily real disposable incomes of households *directly*—by way of tax cuts and increases in transfers and public consumption. The inference that expansionary tax and expenditure measures undertaken to boost disposable incomes of households tend to be large and frequent immediately before elections is, in fact, quite well confirmed by empirical data. For instance, Edward Tufte found in a multi-country study that disposable income usually has tended to go up immediately before elections—in fact in 21 out of 26 democracies studied (Tufte). In many cases, these increases were brought about by fiscal policy measures with a direct impact on real disposable income.

The other macroeconomic variables—inflation, unemployment and the current account—are much more difficult to manipulate by the government. However, let us assume that the government thinks, quite realistically, that it can expand output, at least for a while, by increasing excess demand for commodities and labor, and that the rate of inflation goes up *significantly* only after a time lag. In other words, it is assumed that an expansion of aggregate demand, to begin with, generates a movement up along a given short-term Phillips curve and that an upward shift of the curve will only occur later, if at all. These considerations further strengthen the conclusion that expansionary policy actions should be expected immediately before elections in order to reach a point as close as possible to the origin in a three-dimensional diagram with inflation, unemployment and the current account deficit on the three axes.

The model also predicts that the government will usually cut back aggregate demand after the election in order to bring down the rate of inflation, to squeeze out inflationary expectations and hence to shift down the short-term Phillips curve, as well as to reduce the deficits in the current account, well in time before the next election, so that new expansionary actions can be undertaken again before that election without immediately running into high inflation and a large current account deficit. The experiences of recent years have shown how high the economic and social costs are during this deflationary phase of the policy cycle.

This type of behavior would *contribute to,* though not necessarily generate, "clockwise loops" in a conventional price-Phillips-curve diagram, with the timing, in calendar time, of the loops influenced by the attempts of the authorities to reach positions on these loops as close as possible to the origin during the months immediately before election. Tendencies to such loops can, in fact, be noticed in several countries during the last decade (Grossman; Lindbeck 1975; Nordhaus).

Thus, when the responsibility for macroeconomic stability is relegated to the politicians, it is unavoidable that the interpre-

tation and the implementation of this responsibility become strongly colored by the specific features of the political system, in particular by short-term vote-getting considerations, and that the business cycle therefore becomes a mixture of economic and political forces interacting with each other.

Of course, it is possible to reach more specific conclusions about the timing of these loops if the model is specified in more detail. For instance, it is not only conceivable but likely that the general public is more concerned with the *rate of change* than with the *level* of unemployment, at least as long as the level is not higher than has usually been the case during the postwar period. The reason is that a moderate level of unemployment only hits a small fraction of the population, whereas *increasing* unemployment creates lower job security for a considerable fraction of the population. There is, in fact, some empirical evidence for this in the sense that election outcomes seem to be more sensitive to the *change* than to the *level* of unemployment (Lepper).

If this is a correct hypothesis, and if it is also understood by the politicians, we would expect that the government is satisfied just to have passed "the lower turning point" of the inflation-unemployment loop shortly before an election, in order to reach, before the election, a situation with *falling* unemployment and still relatively modest inflation. On the basis of this theory, it is natural that the governments in the U.S. and West Germany waited so long to take expansionary actions in the 1974-75 stagflation period in view of the fact that the next election in both countries will not occur until the fall of 1976.

Many open economies also show a macroeconomic development pattern where a rapid increase in unit labor costs results for a while in squeezed profits and lagging investment and employment in the tradable sector, as well as a reallocation of resources to the nontradable sector, largely in fact to the public sector. After a while, a devaluation, or possibly a slowing down of wage inflation by way of a more restrictive economic policy, restores profits and investment incentives in the tradable sector, possibly followed by a new phase of falling profits, etc. Thereby a second type of political-economic cycle is superposed upon the short-term business cycle: a profit-investment-devaluation-demand management cycle, usually with a longer periodicity than the ordinary business cycle. We find such a cycle in several countries, such as the United Kingdom, Denmark, Finland, as well as in many LDCs.

### *C. Implications of Endogenous Policy*

*In reality* we would expect that the control by the authorities of the business cycle is very imperfect, judged from the point of view of both the idealistic and the popularity variables. The earlier mentioned loops in inflation-unemployment space would therefore be expected to be rather irregular. There are several reasons for this: (1) the high frequency of exogenous shocks that continuously hit the economy, for instance from international sources; (2) the limited understanding among politicians and their advisers of the functioning of the very complex macroeconomic system; (3) the "imperfections" in the political-administrative system, including the celebrated "policy lags"—recognition lags, decision lags, and effect lags; and (4) the delicate balancing of the interests and opinions of various organizations which challenge the authority of the national government and often succeed in modifying both the actions of governments and the effects of such action.

All this put obvious limits on the possibilities of a successful fine tuning of the mixed economic and political business

cycle.

In the context of this model—with a complicated interaction between markets and policies—it is not downward flexibility of the price and wage *level* that is important for the employment level, but rather the flexibility downwards in the *rate of change* in wages and prices. Moreover, it is not mainly the *automatic* effects of price and wage changes on private behavior, such as on demand for commodities and labor—via real wages, interest rates and real balances—that is important, but the effects of such flexibility on *government behavior*.

More specifically, the more sensitive downwards the rate of price change is to reductions in aggregate demand, the shorter are the periods of unemployment that are necessary. In other words, the lengths of the "stop" periods in the go-stop cycles would be expected to be inversely related to the degree of downward flexibility of the rate of inflation. Hence, it is crucial for economic stability that institutional changes are brought about which make the rate of wage and price increases *more sensitive downward* to reductions in capacity utilization—as well as, of course, *less sensitive upward* to increased capacity utilization. Thus, the lag between quantity and price changes is both a prerequisite for the contemporary "political business cycle" and a reason for the duration of the stagflation periods.

Unfortunately, it is not very clear how such policies and institutional reforms could be brought about. In the case of commodity markets, increased competition, including international competition, and public studies on price setting behavior are often mentioned as potentially useful reforms (though such reforms may increase price flexibility *upward* as well, at very high levels of capacity utilization). Personally I would not object to the breaking up of some large corporations into somewhat smaller units. In the case of the labor market, increased competition, by way of a drastically reduced role of labor unions for wage formation, is hardly a feasible strategy in most countries—in view of the political role of labor unions. However, an increased understanding in society that aggressive wage policies and rigidity of relative wages in fact create and prolong unemployment periods could certainly help. It is perhaps too much to hope that governments would be able to limit inflationary wage increases by *announcing* before wage bargaining that aggregate demand in money terms will only be allowed to increase by a certain limited amount—though the West German government seems to have done just about that on some occasions, with considerable success.

It is very unfortunate that we economists are so empty-handed on this strategic issue in stabilization policy concerning institutional reforms that make the rate of price increase less sensitive upward and more sensitive downward to variations in capacity utilization.

### *D. Endogenous Targets and Instruments*

A changing relative emphasis over time of the various target variables has been analyzed here as movements along a given preference function, with a given set of policy instruments, rather than as shifts in this function. A more general approach would also include as endogenous elements in the analysis (1) shifts in the evaluation of the target variables and (2) the development of new policy tools.[5]

One hypothesis along the first line would be that "policy successes" for a certain

[5] Changes in the tradeoff possibilities can, of course, also have fundamental consequences for such "shifts of emphasis." For instance, governments have often reacted to increased frictional or structural unemployment (with large elements of "voluntary unemployment") by further increasing aggregate demand—also when the frictional or structural changes are in fact caused by the government itself, for instance via more liberal rules for unemployment benefits.

target variable create a tendency among politicians to raise the level of aspiration for this specific policy target. An obvious example is the increasingly ambitious formulations of the "full employment" target in many countries during the 1950's and 1960's when employment policy was very successful. Politicians started to argue that a low general level of employment was not enough, and also that "remaining pockets" of unemployment for specific regions, branches, and types of labor should be removed, thus making the unemployment target more and more differentiated.

Moreover, from the mid-1960's the tolerance for inflation increased in most countries along with the failures to achieve price stability; a rate of inflation characteristic of the peaks of previous booms became more and more regarded as minimum achievable "low points." Similarly, the failures in full employment policy during the international recession of 1974-75 seemed to have lowered dramatically the ambitions of full employment policies; levels that only recently were regarded as "politically impossible" were after a while regarded as levels to strive for, when the observed rates were much higher.

Thus, adjustments of the political aspiration level seem to work in both directions —raising aspirations as a response to success, and reducing aspirations as a response to failures—rather similar to the "psychological" analysis of household behavior by "the Katona school," and the analysis of "satisfying behavior" of firms by "the Carnegie school."

Moreover, both the successes and the failures of stabilization policy seem to stimulate a search for new policy tools. For instance, the early successes of broad full employment policies, and therewith connected increasingly ambitious full employment targets, made politicians develop new policy tools designed to deal with the new more ambitious targets, such as selective subsidies and retraining programs to remove pockets of unemployment in specific regions and branches of industry, and for specific groups of employees such as married women, minority groups, the handicapped, etc. Similarly, the striking policy failures in many countries with respect to price stability and balance-of-payments equilibrium in the early 1970's generated a search for a whole new set of policy tools or the restoration of tools which had not been used for a long time—such as price and wage controls, new types of selective taxes and subsidies, and various tools designed to create barriers between domestic and foreign asset markets.

The general conclusion would be that whereas success and failure with a policy target seem to change the level of aspiration up *or* down in response to successes and failures, respectively, both successes and failures seem to generate a search for new policy tools. The result is a gradual proliferation of new policy tools.[6]

These observations provide additional examples of the importance of endogenizing the behavior of politicians.

## IV. Concluding Remarks

Three basic weaknesses of national stabilization policy have been scrutinized in this lecture. The first one was connected with the international economic influences on the domestic economy, and the second, which is common for open and closed economies, with the inadequacy of aggregate demand management. The suggested remedy for both these weaknesses was to amend aggregate demand management with more differentiated policy tools, mainly designed to influence relative

[6] These generalizations probably have exceptions. After a period of particularly large policy failures, it is conceivable that the policy ambitions are lowered so much that fewer policy tools than earlier are necessary.

prices of commodities and factors as well as profitability, but also to increase factor mobility and to affect the *composition* of aggregate demand and aggregate supply.

However, policies along these lines would further accentuate the third discussed weakness of stabilization policy: that the functioning of the political system is not always in good harmony with the requirements of stabilization policy. The implementation of differentiated and rather sophisticated policy tools would create severe risks of further macroeconomic disturbances from the government—both because of imperfections in the political system and because governments may be more interested in stabilizing votes in the short run than the economy in a somewhat longer perspective. There is also a severe risk of distortions in the allocation of resources by way of deformations of the price system, in particular if policy interventions that are meant to be temporary will in fact be permanent. A heavy concentration of powers in a small group of civil servants and politicians may also arise.

In view of the experiences of economic policy so far, a *pessimist* or a *cynic* may even be tempted to say that the most severe difficulties of economic policy are imbedded in the political rather than in the economic system and that the main obstacle for a successful stabilization policy is, in fact, the government itself. Consequently, the best thing to do would be to avoid discretionary policies altogether rather than trying to make the interventions more sophisticated. He (she) can, of course, support this position by the reasonable assertion that the really dramatic macroeconomic disturbances during the last decade, and perhaps earlier as well, have been the result of government policies, rather than of instabilities in private behavior. In particular, it is difficult to avoid the empirical generalization that inflationary policies followed by restrictive, unemployment-creating actions have been a dominant macroeconomic feature.

*A careful optimist,* by contrast, may perhaps be willing to take some modest steps in the direction of amending aggregate demand management with more differentiated policy tools, for instance in order to influence the real wage rate and to increase factor mobility. A slightly more interventionist strategy would be to try some "semi-selective" management of demand and supply for rather *broad* components of aggregate demand and output —private consumption, fixed business investment, inventory investment, housebuilding, etc.

The *ultra-optimist* may be willing to go all the way to selective demand and supply management for different sectors and geographical regions.

Alternatives to experiments with some selective or semi-selective tools are either to reduce the ambitions in full employment policy somewhat, or to accept high and possibly accelerating inflation. It is, of course, also possible to reduce the "social costs" of macroeconomic instability, for instance by "bribing" firms to hoard "unnecessary" labor during recessions and by bringing about large-scale indexation of credit contracts and tax rates. I am then disregarding an extensive use of direct price and wage controls, which I do not have much faith in.[7]

However, the main message of the lecture has probably been that as we have two interacting systems, the political and the economic, we cannot control one with the other, but we must try to redesign them *both* to improve the stability of each. Most of my suggestions here for such redesigns have referred to the "intersection" of the two systems, as reflected by the policy instruments.

[7] I have no use in this lecture for the unclear concept "incomes policy."

The other suggestions for institutional reforms of the *economic system* have, I am afraid, been more wishful thinking than concrete proposals: to increase factor mobility and the flexibility of relative prices of commodities and factors, *hoping* that this will make both the rate of inflation more sensitive to reduced capacity utilization and less sensitive to increased capacity utilization. The really crucial issue in this context is, in my judgment, to make the labor market less inflationary and to prevent competing claims for income shares (or relative factor returns) from making stabilization policy unmanageable.

It is even more difficult to suggest changes in the *political system* that would improve macroeconomic stability and in particular make politicians follow what perhaps should be the Eleventh Commandment: "Thou shalt not *start* inflating!" However, features that may be singled out for consideration of reform are:

(1) the length of the election period, and the authority of the government to decide the time of election;[8]

(2) the choice between one-party government (or small coalitions) and broad coalitions for the purpose of limiting the "misuse" of short-term stabilization policy as an element in party-political competition;

(3) experiments with "depoliticization" of some policy agencies, such as the central bank;

(4) a more liberal use of discretionary powers of the cabinet without previous consent of parliament;

(5) a more critical attitude by economists, and citizens in general, to attempts by governments to buy votes by way of nonsustainable expansions;

(6) a new assignment of policy instruments and responsibilities between local, national, and international authorities.

The last point is perhaps the most important one. In particular, it is rather clear that the internationalization of the economic system during the post-World War II period has, in principle, considerably reduced the comparative advantage of the national government, relative to international coordination and to supernational organizations, in the field of stabilization policy. A good example is the need to control the supply of international credit, money, and liquidity. Another example is provided by the 1974–75 recession, when all countries seemed to hope for an export-led expansion, generated by expansionary policies in *other* countries!

To summarize, a solution of the problem of macroeconomic stability probably requires not only new and more elaborate policy instruments, as discussed above, but also some redesigning of both the economic and the political system—though, of course, without risking the two basic pillars of Western pluralism: the market system and political democracy. The present "crisis" in the Western economies is not, I think, mainly a "crisis in economics," though we economists have no doubt overestimated the stability of macroeconometric behavior functions. The main problem is not that we are unable to understand analytically what is happening, but rather that the institutional changes and the discretionary policies that are necessary for macroeconomic stability seem to be politically difficult to implement.

[8] For instance, a theoretical possibility would be to let the exact time of elections be determined by a random process.

## REFERENCES


J. Åkerman, "Political Economic Cycles," *Kyklos*, 1947, *1*, 107–17.

O. Aukrust, "Inflation in the Open Economy," The Central Bureau of Statistics of Norway, mimeo, 1974.

L. Calmfors and E. Lundberg, *Inflation och*

*arbetslöshet*, Stockholm 1974.

A. **Downs**, "An Economic Theory of Political Action in a Democracy," *J. Polit. Econ.*, Apr. 1957, *65*, 135–50.

G. **Edgren**, K. O. **Faxén** and C. E. **Odhner**, *Wage Formation and the Economy*, London 1973.

B. **Frey** and L. J. **Lau**, "Towards a Mathematical Model of Government Behaviour," *Zeitschrift für Nationalökonomie*, Dec. 1968, *28*, 355–80.

H. I. **Grossman**, "The Cyclical Pattern of Unemployment and Wage Inflation," *Economica*, Nov. 1974, *41*, 403–13.

P. J. K. **Kouri**, "The Exchange Rate and the Balance of Payments in the Short Run and in the Long Run: A Monetary Approach," *Scandinavian J. Econ.*, 1976, *78*, No. 2.

G. H. **Kramer**, "Short-Term Fluctuations in U.S. Voting Behavior, 1896–1964," *Amer. Polit. Science Rev.*, Mar. 1971, *65*, 131–43.

S. J. **Lepper**, "Voting Behavior and Aggregate Policy Targets," *Public Choice*, Summer 1974, *18*, 67–81.

A. **Lindbeck**, *A Study in Monetary Analysis*, Stockholm 1963.

———, "Endogenous Politicians and the Theory of Economic Policy," Seminar pap. no. 35, Institute for International Economic Studies, Univ. of Stockholm 1973.

———, "Is Stabilization Policy Possible? Time-lags and Conflicts of Goals," in W. L. Smith and J. M. Culbertson, eds., *Public Finance and Stabilization Policy*, Amsterdam 1974, 269–307. (Also Reprint No. 34, Institute for International Economic Studies, Univ. of Stockholm.)

———, "Business Cycles, Politics and International Economic Dependence," *Skandinaviska Enskilda Banken Quarterly Rev.*, 1975, No. 2, 53–68 (Also Reprint No. 39, Institute for International Economic Studies, Univ. of Stockholm.)

W. D. **Nordhaus**; "The Political Business Cycle," *Rev. Econ. Studies*, Apr. 1975, *42*, 169–90.

G. **Rehn**, "Some Thoughts on Employment Policy in the OECD Area during 1975 and Later Years," mimeo, 1975.

E. R. **Tufte**, "The Political Manipulation of the Economy: Influence of the Electoral Cycle on Macroeconomic Performance and Policy," Department of Politics, Princeton Univ., mimeo, Sept. 1974.

# Marx and the Falling Rate of Profit

*By* JENS CHRISTIANSEN*

The renewed interest in Marxist economic theory during the past few years brought with it a renewed interest in what Marx himself called "the most important law of modern political economy," the "Law of the Tendency of the Rate of Profit to Fall."

K. Marx's own analysis of the laws of motion of capitalist development led him to the following formulation of this law. In the course of capitalist development there exists a tendency for live labor to be continuously replaced by dead labor. Capitalists replace workers by using more machinery. At the same time the increased productivity of the labor process lead to an increased amount of raw material and auxiliary material processed in any given hour of labor. Thus the amount of dead labor incorporated in the means of production (machinery, raw material, and auxiliary material) increases over time relative to the amount of live labor directly employed in the production process. Marx expresses this development in what he calls the rising organic composition of capital, i.e., an increasing ratio of dead labor to live labor. This development, Marx argued, would lead to a tendency of the rate of profit to fall over time. Thus, for him, this law is not a prediction at the level of appearances. Rather he claims that despite the existence of a number of counteracting forces, the tendency of the rate of profit to fall would assert itself as a continuous structural threat to the capitalist mode of production. On the one hand Marx clearly saw this tendency as a long-run development of the capitalist mode of production. It is, in this sense, directly related to the inherent dynamics of the process of capital accumulation. On the other hand he put great emphasis on linking up this tendency with short-run crisis phenomena. Marx saw crises as temporary breakdowns in the capitalist accumulation process. Yet at the same time he considered them to be a mechanism that recreates the preconditions for profitable capital accumulation. Crises are an inherent part of the continued reproduction of capitalism. "In the totality of this disorderly movement is to be found its order" (Marx 1933). One effect of crises is to reduce the rate at which real wages rise or even to cause a decline in real wages. Furthermore crises result in the devaluation of capital as well as its concentration and centralization. It is precisely these effects which allow the accumulation process to continue by reestablishing profitable investment opportunities.

Marx's own formulation of the law

* Stanford University. This paper is a byproduct of ongoing research with Bill Lunt which originated in the Political Economy Seminar at Stanford. Furthermore, I greatly benefited from comments on an earlier draft by Sue Bessmer, Michael Carter, Duncan Foley, Don Harris, Bridget O'Laughlin, Anwar Shaikh, Sandy Thompson, and Robert Williams. Unfortunately, because of limitations in time and space, I was unable to incorporate all useful comments and suggestions.

never took final form and contains a number of ambiguities. The different forces that influence the development of the organic composition of capital are never systematically put together. Nor is the relationship between the organic composition of capital and the rate of profit fully worked out. This situation has led to an ongoing debate over what the correct interpretation and application of the law is. Many of the contentions in this debate have led Marxists into wrong directions. At the empirical level the debate has centered primarily around the question of whether or not the organic composition of capital is indeed rising. At the theoretical level the argument has been predominantly concerned with the causes of a rising organic composition and the question of whether a rising organic composition will necessarily lead to a falling rate of profit. There seems to be enough progress on some fundamental points, however, so that we can eliminate much of the confusion that underlies many of these earlier debates.

The present paper will briefly deal with some of the criticisms of the law and with some of the inadequacies in alternative formulations that have been proposed. I shall try to show, in a very schematic way, that none of the existing formulations of the law have been fully adequate in what they are trying to do, viz. to lay bare the laws of motion of the capitalist mode of production. The second part of the paper will lay out a programmatic scheme for dealing with the issues. This scheme will be based on two aspects that Marx clearly identified in treating the falling rate of profit. The first relates to long-run trends in the process of capital accumulation. The second concerns the role of crises in the movement of the rate of profit.

## I. Criticisms and Alternative Formulations

In dealing with the existing Marxist literature I shall start with the question of whether or not the organic composition of capital is rising. Two extreme approaches have been taken. On the one hand it has been treated as a question the answer to which can be theoretically deduced from the definition of capital as self-expanding value.[1] This is clearly inadequate because it claims to understand existing material reality on the basis of idealist concepts created by the human mind. Such an approach is incapable of understanding changing historical structures.

On the other hand the question has been treated as a purely empirical problem. This is incorrect because the law is conceived at the abstract level of the capitalist mode of production. Thus all the links between this level of abstraction and the level of concrete reality have to be theoretically specified before an empirical investigation makes any sense. This involves questions of the accurate theoretical definition of the categories involved, their appropriate unit of analysis (e.g., national or world economy), and their historical specificity. The existing empirical studies fall short of most of these criteria. It is therefore not surprising that their results are contradictory and inconclusive. J. Gillman purports to show that the organic composition of capital for the *U.S.* economy has not shown any sustained tendency to rise since about 1920. Many Marxists and even more non-Marxist economists have shared that view. Victor Perlo, on the other hand, finds that the *U.S.* capital-labor ratio rises between 1929 and 1963. Shane Mage's results go into the same direction. However, he claims that the observed drastic fall "in the Marxian rate of profit for the U.S." is based to a major degree "on a substantial long-term decline in the rate of exploitation" (Mage, p. 3). A theoretical issue which also underlies the

[1] Some of the statements by David Yaffe and Mario Cogoy would fall into this characterization.

above empirical studies involves the definition of the organic composition of capital and its relationship to the rate of profit. Most Marxists have interpreted the organic composition of capital to be the ratio of constant to variable capital. The rate of profit (in value terms) is then a function of this ratio (the value composition of capital) and the rate of surplus value (the ratio of surplus value to variable capital). Both Joan Robinson and Paul Sweezy have pointed out that even if the value composition of capital increases with capitalist development, there is no reason to assume that this tendency cannot and will not be counteracted by a continuous increase in the rate of surplus value. Thus the development of the rate of profit over time is indeterminate, despite a rise in the value composition of capital.

This criticism is correct as long as we accept the interpretation that the value composition and the organic composition of capital are the same ratio. Once we take the organic composition to be the ratio of dead to live labor, then Marx's statement of the law is logically correct. This is easy to see. All we have to show is that the organic composition of capital is the inverse of the maximum rate of profit. In the case of zero wages the rate of profit becomes the ratio of surplus value to constant capital and the organic composition of capital becomes the ratio of dead labor (constant capital) to surplus value. Thus, if the organic composition of capital rises without limit, the rate of profit will eventually fall, no matter how much the rate of surplus value is going to rise.

This formulation of the law[2]—though logically correct—does not specify the actual mechanism by which changes in the organic composition lead to changes in the rate of profit.[3] A mathematical relationship that holds in the limit does not, in and of itself, imply any casual relationship between the two categories. Furthermore—as it stands—this formulation does not deal with the question of why the organic composition should rise. Some theoretical studies have tried to deal with these questions by analyzing why capitalists would introduce techniques with a higher organic composition of capital which could, at least potentially, lead to a fall in the rate of profit.[4] No unambiguous answer has been found. The reason for this is quite simple. It lies in a faulty methodology in these approaches. It is assumed that technical change comes about through capitalists choosing, on the basis of the profit criterion, among different existing techniques of production with a higher or lower organic composition of capital. In other words, the basic approach assumes given choice situations for capitalists. It thus neglects the more fundamental question of how these choice situations are generated by the historical development of social forces.[5] The whole point of Marxist social theory is to understand the determinants of these social forces, rather than to make predictions about the real world on the basis of mathematical relationships. The former can only be achieved through a concrete historical materialist analysis of the capitalist accumulation process based on an adequate theoretical conceptualization of the categories and relationships involved.

Before we turn to the issues concerning such analysis, one further criticism of the law must be considered. A correct treat-

[2] See Cogoy, Yaffe and Heinz Holländer.

[3] See Susan Himmelweit.

[4] See Himmelweit, Holländer, Nobuo Okishio.

[5] Exactly the same criticism applies to the voluminous bourgeois literature on the question of whether technical change has a labor-saving bias or capital-saving bias. For an interesting critique of this literature, see Nathan Rosenberg 1969.

ment of the transformation problem (between labor values and prices of production) shows us that in general the rate of profit in value terms does not equal the rate of profit in price terms (prices of production).[6] We are, however, interested in the latter rate because it is precisely in the prices of production scheme that an equal rate of profit for the economy as a whole gets established. Thus the appropriate conceptualization of the law has to be based on a prices of production scheme, with the overall organic composition of capital being a weighted average of the organic compositions of the different sectors, the weights corresponding to the proportions of direct labor used.[7]

## II. Accumulation and Crisis and the Rate of Profit

Where does all this leave Marx's law of the tendency of the rate of profit to fall? It has become clear that neither Marx's own formulation nor any of the subsequent approaches have been fully adequate. The two issues that have up to now dominated the literature are the question of whether or not the organic composition of capital is rising and secondly the causes for an increasing organic composition and its relation to the movement of the rate of profit. These issues are clearly central to the law of the tendency of the rate of profit to fall. However, the above critical evaluation of the literature has shown that these issues can only be understood within a framework that consists of two separate but closely interrelated parts. The first concerns the long-run development of the rate of profit within the capitalist accumulation process. The second involves influences of short-run crises on the movement of the rate of profit.

[6] See Ian Steedman, Alfredo Medio, and Christiansen and William Lunt.

[7] See Christiansen and Lunt.

Concerning the first part, there is a fundamental lesson that we can learn from a careful reading of Marx. An understanding of the long-run development of the organic composition of capital and of the rate of surplus value can only come out of a concrete historical analysis of the nature of technology and technical change under capitalism. That requires a careful analysis of the labor process and the changes it has undergone in the continuous process of concentration and centralization of capital which are integral parts of the process of capital accumulation. For Marx this process of accumulation is the particular form in which the dialectical interaction between forces and relations of production manifests itself under capitalism. Technical relations between people and nature, i.e., forces of production, always imply corresponding forms of social relations among people. Therefore we cannot analyze these two in isolation from each other. Any changes in technology will change the way people relate to each other and vice versa.

On the one hand there are many examples of how "complex technologies create internal compulsions and pressures which, in turn, initiate exploratory activity in particular directions" (Rosenberg 1969, p. 4). Technical imbalances between interdependent processes of production create strong stimuli for innovations. One example of such a technical imbalance would be Bessemer's development of a new process of steel production. Bessemer's

> . . . attention was directed to the subject of artillery by the outbreak of the Crimean War in 1854. He developed a new gun of superior power, capable of firing a heavy, elongated projectile. To his dismay, however, he concluded that the gun could not be safely fired if it were constructed of the standard cast iron employed in artillery pieces. The Bessemer process was the product of Bessemer's search for a super-

> ior metal which would make his newly designed gun practicable. [Rosenberg 1969, p. 7.]

Despite the fact that the initial push for a specific technical change in this example resulted from technical conditions, it should be stressed that such technical change will in general only come about and be maintained if it is compatible with the existing social relations, e.g., if it is profitable.

On the other hand there exist innumerable examples of technical change resulting in the first instance from the dynamics of social relations, the most drastic case in point being technical innovations initiated by capitalists as a response to working class threats against capitalist relations of production. The threat of the withdrawal of labor services through strikes has again and again been "a powerful force in directing energies in search for labor-saving machines" (Rosenberg 1969, p. 13). To provide just a few examples:

> The self-acting mule, the greatest invention of modern industry, put out of action the spinners who were in revolt. [Marx, n.d., p. 161]

> The invention of making gun barrels by means of grooved rolls is due to a Birmingham manufacturer of the name of Osborne. It was on the occasion of a strike of the barrel welders that he was led to make the experiment. [Goodman 1866, p. 389]

> The uncertainties which the supply of labor posed in the British gun-making industry led to the first large scale borrowing of American technology by the Old World. [Rosenberg 1969, p. 15]

Once again, despite the initial push this time coming from the social relations of production, these inventions were possible only because technology had reached a certain point prior to the invention.

Many of the above examples would suggest a development of technology which increases the technical composition of capital, i.e., the ratio of means of production (including in particular raw materials) in physical terms to the amount of labor employed. It is obvious, however, that a much more systematic historical and theoretical analysis is necessary. Within such an analysis one would also have to examine the crucial question of the development of productivity over time. A rising technical composition of capital does not necessarily imply an increase in the organic composition of capital. A careful reading of Chapters IV and V of the third volume of *Capital* clearly indicates that Marx himself was fully aware of the importance of certain trends that would lead to a decrease in the organic composition of capital. Such trends would include, for example, all measures that reduce the turnover period of capital.

Furthermore, we have to pay careful attention to the development of real wages as the third factor—besides the technical composition of capital and productivity—determining the organic composition of capital.

Marx himself attached special importance to the role of the capital goods sector in the generation of technological change.[8] Economies of scale are of particular significance in this respect.

H. Braverman's *Labor and Monopoly Capital* can serve as an important source for developing the necessary analysis of the labor process in the monopoly stage of capitalist development.

The second part of our analysis concerning the law of the tendency of the rate of profit to fall can also benefit greatly from a careful reading of Marx. In Chapter XV of the third volume of *Capital*, Marx emphasizes that the development of

[8] See Rosenberg 1976.

the rate of profit under capitalism is integrally linked to and cannot be understood separately from crises that interrupt and simultaneously recreate the conditions for the process of capital accumulation.

The accumulation process that I have talked about above is certainly not seen as a smooth process by Marx. Its dynamics come directly out of the basic contradiction between capital and labor in their struggle over the control of the production process and the appropriation of the final product.

This struggle will at certain times lead to the closing off of profitable investment opportunities at least for the weaker capitalists. At such a point spending of capitalists and consequently of workers declines. Thus the system as a whole will be confronted with insufficient demand. Not all commodities produced can be sold and the crisis of accumulation manifests itself as a realization crisis. This will lead to bankruptcy and unemployment. The consequence will be further concentration and centralization of capital and simultaneously a slow-down or even a decline in real wages. Inefficient capital will get destroyed in the process. All of these developments will eventually recreate the conditions for profitable capital accumulation by raising the rate of exploitation and lowering the organic composition of capital.

This outline is indeed extremely sketchy. Serious work remains to be done on the Marxist theory of crisis. In particular the role of the money and credit system within the crisis is in need of further systematic analysis.

### III. Conclusion

The above very sketchy review of the literature on Marx's law of the tendency of the rate of profit to fall has shown us that—at present—no satisfactory formulation of this law exists. However, Marx's own ideas on this subject still seem to provide us with extremely useful insights into the capitalist accumulation process and its concomitant crises.

### REFERENCES

**H. Braverman,** *Labor and Monopoly Capital,* New York 1974.

**J. Christiansen and B. Lunt,** "Marx's Law of the Tendency of the Rate of Profit to Fall," unpublished 1976.

**M. Cogoy,** "Les Théories Neo-Marxistes, Marx et l'Accumulation du Capital," *Les Temps Modernes,* Sept.-Oct. 1973, 396–427.

———, "The Fall in the Rate of Profit and the Theory of Accumulation," *Bulletin of the Conference of Socialist Economists,* Winter 1973, 52–67.

**J. Gillman,** *The Falling Rate of Profit,* New York 1958.

**J. D. Goodman,** "The Small Gun Trade," *Birmingham and the Midland Hardware District,* London 1866.

**S. Himmelweit,** "The Continuing Saga of the Falling Rate of Profit—A Reply to Mario Cogoy," *Bulletin of the Conference of Socialist Economists,* Autumn 1974.

**H. Holländer,** "Das Gesetz des tendenziellen Falls der Profitrate, Marxens Begründung und ihre Implikationen," *Mehrwert,* June 1974, *6,* 105–135.

**S. Mage,** *The Law of the Falling Tendency of the Rate of Profit,* unpublished Ph.D. dissertation, Columbia Univ. 1963.

**K. Marx,** *Capital, 1, 2, 3,* New York 1967.

———, *Wage-Labour and Capital,* New York 1933.

———, *The Poverty of Philosophy,* Moscow n.d.

**A. Medio,** "Profits and Surplus-Value: Appearance and Reality in Capitalist Production," in E. Hunt and J. Schwartz, *A Critique of Economic Theory,* Harmondsworth 1972.

**N. Okishio,** "Technical Change and the Rate of Profit," *Kobe University Economic Review,* 1961, *7,* 85–99.

**J. Robinson,** *An Essay on Marxian Economics,* London 1966.

**N. Rosenberg,** "The Direction of Techno-

logical Change: Inducement Mechanisms and Focusing Devices," *Economic Development and Cultural Change*, Oct. 1969, 1–24.
———, "Marx as a Student of Technology," *Monthly Review*, forthcoming.
A. **Shaikh,** "Marx's Theory of Value and the So-Called 'Transformation-Problem,'" unpublished 1974.
I. **Steedman,** "Value, Price and Profit," *New Left Review*, Mar.-Apr. 1975, *90*, 71–80.
P. **Sweezy,** *The Theory of Capitalist Development*, New York 1942.
———, "Some Problems in the Theory of Capital Accumulation," *Monthly Review*, May 1974, 38–55.
D. **Yaffe,** "The Marxian Theory of Crisis, Capital and the State," *Economy and Society*, May 1973, 186–232.

# Stagnation, Instability and International Competition

*By* RAFORD BODDY AND JAMES R. CROTTY*

Significant economic events typically find their reflection in theoretical controversies. The crisis of *U.S.* and world capitalism is no exception. Over the last several years there has been a resurgence of interest in Marxian crisis theory, that part of Marxian theory which deals with the inevitable impediments within capitalism to prolonged growth. Obviously, crisis, depression, and recovery are integrally related. In this essay, we are particularly concerned about the nature of recovery because we believe that a correct understanding of its dynamics is central to the development of a correct theory of the laws of motion of modern capitalism.

For our purposes, theories of the recovery can be divided initially into those which imply that endogenous forces unleashed in the crisis and depression eventually generate a strong recovery and those which deny that such endogenous forces exist. Associated with the former group would be such business cycle theorists as Arthur F. Burns and Wesley C. Mitchell as well as those Marxists, including ourselves, who believe that the depression eventually recreates a potential for profit-making sufficiently strong to renew the accumulation process. The latter group would seem to include neo-Keynesian disequilibrium theorists such as Robert Clower and Axel Leijonhufvud who have critically analyzed unemployment but have as little to say about the way in which the economy gets out of a crisis as they do about how it gets in. However, the most important set of theorists who deny the possibility of endogenous recovery are the stagnationists. In the Marxist tradition, stagnation theorists have been extremely influential. They include Michal Kalecki, Paul Baran and Paul Sweezy—justifiably ranked among the most eminent political economists of the modern period—and J. Steindl whose seminal work of 1952 is again in press.

Because it has provided an explanation of the workings of modern capitalism consistent with such important developments as the Great Depression, the immense growth of the state sector, and the high unemployment of the late 1950's and early 1960's, the monopoly-based neo-Marxist theory of stagnation is undoubtedly attractive. But, while the theory provides useful insights into the historical process by which monopolization creates certain impediments to expansion and exacerbates the dislocations of depression, its denial of the long-run recovery potential of modern American capitalism is seriously misleading.

In this exploratory essay we briefly review the tenets of stagnation theory, point to some of its main weaknesses, and examine its consistency with post-World War II economic developments. To anticipate the principal conclusion, we find that secular changes of the productive forces in the international sector have led to increasing competition and intercapitalist rivalry and thereby have gradually re-

* American University and University of Massachusetts at Amherst. Data references have been omitted due to space limitations but are available on request.

duced the explanatory power of the theory of chronic stagnation.

## I. Tenets and Problems

The central proposition of the most influential variant of neo-Marxian stagnation theory is that the rising surplus in the monopolized core of the economy cannot be absorbed by the private sector. (On the concept of the surplus, see Baran and Sweezy, pp. 9–10.) In the absence of increased state sector stimulation (dramatically evident in wartime) or exogenous, epoch-making innovations, such as the growth of the railroads or the auto industry and its derivative sectors, monopoly capital, it is argued, should be characterized by high and rising unemployment rates, decreasing rates of capacity utilization, ever weaker expansions, longer contractions, and a rise in the percentage of resources devoted to wasteful, nonproductive activities such as the sales effort. Correspondingly, given strong exogenous stimulation, monopoly capital should exhibit dramatically rising levels of profit and cash-flow, an increasing rate of profit, and a corresponding decline in labor's share of income. Since the state has accepted responsibility for overall economic performance, there should be a tendency for the state sector to grow relative to the rest of the economy.

Among Marxists, the most strongly questioned part of the theory has been the tendency of the surplus to rise. This is natural enough since Baran and Sweezy themselves note that the law immediately invites comparison with the classical Marxian law of the falling tendency of the rate of profit. Data on markups at given levels of capacity utilization presented by Baran and Sweezy (p. 84) support their formulation. On the other hand, more recent data on the same series presented by J. Blair (p. 641) provides evidence to the contrary. A corollary of the rising surplus is the ability of the corporation to raise prices during periods of high demand by an amount sufficient to offset or exceed any increase in unit labor costs. As evidenced in the data of the postwar period, however, labor has been able to increase its share in the latter part of such periods (see Boddy and Crotty). This evidence casts doubt on the alleged tendency of the surplus to rise.

We are equally concerned with the validity of the stagnationists conclusion that modern capitalism has no significant long-run endogenous outlets for that surplus. As the stagnationists see it, the fierce drive of the capitalist to accumulate and grow is in no way diminished by oligopolization. Yet, the surplus is presumed to be effectively bottled up by the decline in profitable outlets. The corporation, it is argued, does not want to spoil its own market through expansion and is hesitant about investing outside of its industry. However, it would appear that there are many channels whereby any excess surplus can and does find outlets. The surplus, as noted by the stagnationists, can be thrown into the sales effort. The corporation can integrate vertically. It can invade other product markets, particularly in the competitive sectors, directly or through conglomerate mergers. It may expand geographically within the country or, as in recent decades, internationally. Indeed, each of these outlets has been cited by Sweezy himself. And, as monopoly firms pour capital and labor into the competitive sector, the ensuing expansion feeds back into the monopoly sectors. Moreover, the financial sector, though far from frictionless, does move surplus to areas with profitable investment opportunities. Each of these outlets has been aggressively utilized by the corporations in the postwar period.

The fact that *U.S.* and international capitalism have been characterized by

substantial growth in the last half century has obviously not escaped the notice of such astute observers as Kalecki, Baran and Sweezy. Rather, the growth has been primarily attributed to wars and epoch-making innovations assumed to be exogenous. But, if the law of capitalism remains "accumulate, accumulate," then there is a perpetual tendency to revolutionize the means of production, and pressures toward path-breaking innovations are built into the system.

The question of what is and what is not to be taken as exogenous is, of course, a long-standing problem. For Marxists, even the surplus absorption of wars has a strong endogenous aspect. As Myron Sharpe has aptly remarked, "Why can we not argue that the main tendency of monopoly capital is major innovations and wars, these being interrupted at times by abnormal periods of stagnation?" (P. 46.) The major point we wish to emphasize is that monopoly capitalism is still capitalism. The drive to grow, to accumulate, to take advantage of and *create* profit-making opportunities is not dulled. While the growth of oligopoly has created the likelihood of more severe crises and depressions as explained by the stagnationists, capitalism has not been shorn of its endogenous expansionary capabilities.

## II. Theory and Reality: The Mid-1950's Through the Early 1960's

For a significant period of time during the post-World War II era, particularly from 1956–57 through 1962–63, the central deductions of stagnation theory were in large part consistent with the empirical evidence. Corporations were able to finance most if not all of their expenditures on physical assets out of internally generated funds. Expenditures on the sales effort and planned obsolescence were clearly rising. Moreover, the economy could be sensibly described as stagnant. From 1958 through 1963, the rate of unemployment averaged almost 6.0 percent, fluctuating from an annual low of 5.5 percent to a high of 6.8 percent. From 1956 through 1961, the average annual rate of growth of real *GNP* was only slightly in excess of 2 percent.

The stagnationists associate a large government sector with the absence of depression and a sluggish economy with the absence of active government stimulation. To what extent can these particular conditions of stagnation be attributed to the alleged need for greater and greater government stimulus in the face of an increasing stagnationist tendency? In the context of this question, it is important to distinguish between the failure of the government to positively stimulate the economy—a "neutral" policy—and a policy of active and deliberate restraint. In fact, government policy in this period falls into the latter category. Real purchases of goods and services by federal, state and local governments dropped substantially in 1954 and moderately in 1955. In turn, they were almost constant from 1958 to 1960. At the federal level, where spending is associated with conscious macropolicy objectives, real purchases were lower in 1961 than in 1954. The high employment budget had been in deficit from the third quarter of 1951 through the second quarter of 1954, and had obviously contributed to the boom of that period. However, it was in surplus from the third quarter of 1954 through the third quarter of 1965, with an especially large surplus in 1956–57 and 1960–61. As to monetary policy, we note that total member bank reserves were lower in 1963 than they had been in 1952. Within the period 1954 through 1960, we find that only in 1958, in the midst of a severe contraction brought on in part by restrictive monetary policy, did the monetary base expand at a rate appreciably greater than 1.5 percent per annum.

Why did the policy makers feel compelled to hold down a supposedly stagnating economy? We believe that the evidence is consistent with the hypothesis that macropolicy from the mid-1950's to the early 1960's was kept firmly restrictive in pursuit of two objectives, both of which would have been jeopardized by the normal process of endogenous economic expansion in the absence of such governmental restriction. One was a desire to prevent the inevitable "excesses" endemic in capitalist expansions, such as rising wages, labor militancy, and inflation, from eroding profits, thereby creating the preconditions for a major contraction. The other was a desire to protect the crucial international institutions under *U.S.* control as well as the privileged reserve currency status of the *U.S. dollar* from those pressures generated by sustained expansion which it is clear in retrospect could have led to a rapid decline in the international position of the United States. That macropolicy was deliberately kept restrictive in order to achieve such objectives is made clear in the studies by G. Bach and H. Stein of the history of macropolicy making in this era. The sluggish behavior of the economy during the period 1956–63 therefore cannot be taken as prima facie evidence of the lack of an expansionary dynamic of the private sector.

## III. Theory and Reality: The Last Decade

Whatever attraction stagnation theories may have as an explanation of the experience of the 1950's and early 1960's, an examination of the record of the past decade casts doubts on their long-run adequacy. Provided that sufficient growth in aggregate demand is maintained, it follows in the logic of the theory that the industrial corporate sector should attain rising profits, a growing profit share, and an increasing flow of internally generated funds. Under the expansionary macropolicy of the early 1960's and the military spending of 1965–68, aggregate demand was sustained at a high and increasing level for most of a decade. Yet, starting in 1965 corporations experienced falling rates of profit and were increasingly forced into the credit market to finance accumulation. In nominal terms, internally generated funds of nonfarm, nonfinancial corporations were virtually stagnant from 1966 through 1969, while in real terms they declined substantially. Corporate indebtedness grew dramatically. Now, while the accumulation of consumer and government debt fits comfortably into monopoly-based stagnation theories, *eroding profit margins and accelerating corporate debt in periods of high demand do not.* In like manner, stagnation theory seems singularly unfruitful as a framework within which to analyze the changes which have occurred in the relationship between banks and major nonfinancial corporations. The increasing dependence of major corporations on banks for their financing, well documented by H. Magdoff and Sweezy, seems to move us beyond the confines of stagnation theory.

## IV. The Rise of International Competition

To fully understand the partial empirical success of stagnation theory in the 1950's and its inadequacies during the last decade, we believe that one must include the international sector in the analysis. For this sector has inevitably become an arena of increasing capitalist contention, competition and rivalry. As is now becoming universally acknowledged, profound changes have occurred in the international system in the past 25 years. Parts of the world have been forcibly removed from the direct control of the major capitalist powers. The financial arrangements of Bretton Woods, which included the privilege of *U.S.* seigniorage, have broken

down. There has been an as yet unreversed decline in *U.S.* control over formerly cheap supplies of energy and raw materials. But the most important secular change has been the end of the monopoly power of the United States within the international capitalist system and the concomitant rise of Japan and the European Economic Community countries, especially West Germany, as serious economic and political rivals of the United States. The inability of the United States to suppress conflict and to enforce the international "rules of the game," originally written to maximize its own advantage, has fractured the stability of the international capitalist system.

Over the last fifteen years, the resurgent and technologically advancing industries of Japan and West Germany have subjected *U.S.* corporations to increasingly tough competition in international trade, investment, and the struggle for control of raw materials. From 1960 to 1972, exports in constant dollars grew at 6.0 percent for the United States, 9.7 percent for West Germany and France, and 16.7 percent for Japan. Of course, these trends accelerated appreciably during the Vietnam War. Lest this all be dismissed as an exercise in exponents, we point out that by the terminal year the value of Japanese and West German exports were 55 percent and 95 percent as large, respectively, as those of the United States. The decline in American power and the dilemma it presents for American corporations and policy makers have been succinctly set forth by C. Fred Bergsten in a recent issue of *Foreign Affairs*:

> Not long ago the United States was virtually the sole source of large chunks of investment capital, technological know-how and marketing skills. Not long ago, security considerations devolving from the cold war enabled the United States to dominate its allies, most small countries and the international rules and institutions which governed world economic relations. Now, however, . . . (i)ncreasing numbers of European and Japanese companies offer formidable competition for American multinationals . . . With the onset of detente, host countries, large and small, no longer fear to cross the United States by challenging U.S. based firms and the economic environment that helped them flourish in the 1950's and 1960's. [p. 138]

In this essay we are concerned with the implication of these changes in the international sector on the dynamics of the *U.S.* economy, especially their impact on its potential tendency toward stagnation. The inevitable resurgence of international competition raises the following question: *Will an economy characterized by a tendency toward stagnation because of the oligopolization of its major industries develop serious countertendencies as that economy becomes enmeshed in increasingly severe international competition?* Certainly increasing competition for raw material and markets implies that costs may not be so easily reduced, nor profit margins so easily protected. In like manner, if *U.S.* corporations fail to invest and expand in order to avoid spoiling the market, others will encroach on their market share. International competition means pressure to trim down corporate fat and muscle-up the production process. Significantly, many of the industries, such as autos and steel, which epitomized the oligopolistic market structure at the heart of stagnation theory are now those under most intense international attack.

Many of the "confusing" phenomena of the last decade are more easily understood once we take increasing international competition into account. Increasing competition helps to explain why prices failed to rise sufficiently to cover the increased unit labor costs of the late 1960's. This, in turn, helps explain the precipitous

decline in profit margins and the increasing need for external finance. There are also a number of important recent developments in the realm of government policy making which are quite consistent with a resurgence of international competition. The concern with "capital shortage" and the drive to increase the percentage of *U.S.* output channeled into investment up to the level achieved by *U.S.* competitors, the attempt to reverse the decade-long decline in labor productivity, the pressure to halt or reverse the growth of public employment and welfare programs, and the increased interest in economic planning all seem to be more easily associated with increasing competition than with a perspective of chronic stagnation.

## V. Conclusion

Rooted in the reality of the growth of the giant corporation and the continuous rise of monopoly power, neo-Marxist theories of stagnation have raised valid and interesting questions about the development of additional impediments to sustained expansion and, perhaps more importantly, about the increased potential severity of crises and depressions. Our quarrel is with the concept of chronic stagnation and its denial of the prospect of eventual recovery and subsequent expansion.

We agree with Baran, Kalecki, Steindl and Sweezy that depressions have a tendency, in the absence of state intervention, to be more prolonged in a system of monopoly capitalism. But for that very reason, monopoly capitalism has a tendency to generate even more powerful bursts of expansion and accumulation. In turn, it is the very strength of the accumulation process which creates those excesses endemic in capitalism which eventually rupture the expansion, making crises inevitable and recessions and depressions functional. It is the potential for overexpansion which had led macropolicy-makers to restrain the economy as often as they have stimulated it. Currently, in spite of the extremely high levels of unemployment and unused capacity, macro policy-makers are determined to prevent the mild expansion we are experiencing from generating too much momentum. (For an analysis of the macropolicies of the past decade and a discussion of policy prospects, see Crotty and Rapping.)

Thus, the question at issue is not whether the working people of America, under existing capitalist institutions, can avoid the unemployment, insecurity, and waste of recurring recessions and depressions. The question is the underlying mechanics of crisis and recovery, in particular the very possibility of the latter. We reject the concept of perpetual depression as the natural state of monopoly capitalism, and do not look to monopoly based stagnation theory for the key to understand the origins of the recent crisis of *United States* and world capitalism.

## REFERENCES


**G. Bach,** *Making Monetary and Fiscal Policy,* Washington 1971.

**P. Baran and P. Sweezy,** *Monopoly Capital,* New York 1966.

**C. F. Bergsten,** "Coming Investment Wars?" *Foreign Affairs,* Oct. 1974, 135–52.

**J. Blair,** *Economic Concentration,* New York 1972.

**R. Boddy and J. Crotty,** "Class Conflict and Macro Policy: The Political Business Cycle," *Rev. Radical Pol. Econ.,* Apr. 1975, 1–19.

**R. Clower,** "The Keynesian Counter-Revolution: A Theoretical Appraisal," in R. Clower, ed., *Monetary Theory,* Baltimore 1970, 270–97.

**J. Crotty and L. Rapping,** "The 1975 Report of the President's Council of Economic Advisors: A Radical Critique," *Amer. Econ. Rev.,* Dec. 1975, *65,* 791–811.

M. Kalecki, *Selected Essays on the Dynamics of the Capitalist System,* Cambridge 1971.

A. Leijonhufvud, *On Keynesian Economics and the Economics of Keynes,* New York 1968.

H. Magdoff and P. Sweezy, "Banks: Skating on Thin Ice," *Monthly Review,* Feb. 1975, *26,* 1–21.

M. Sharpe, "Marxism and Monopoly Capital: A Symposium," *Science and Society,* Fall 1966, 461–69.

H. Stein, *The Fiscal Revolution in America,* Chicago 1969.

J. Steindl, *Maturity and Stagnation in American Capitalism,* Oxford 1952.

P. Sweezy, "On the Theory of Monopoly Capital," *Monthly Review,* Apr. 1972, *23,* 1–24.

# International Crisis and Politicization of Economic Activity

*By* Daniel Luria and Arthur MacEwan*

The crisis in the international operation of *U.S.* business has developed from basic changes in world power relations. The decline in the *U.S.* trade position, the breakup of the capitalist world's monetary system, the formulation of new policies toward the Soviet Union and China, the rising strength of the major oil-producing nations, and the final defeat of *U.S.* aggression in Indochina constitute a chronicle of the changing world order.

The international crisis is working in combination with domestic events to produce a far-reaching impact on the nature of economic, social, and political relations within the United States. In this paper, we focus on two aspects of that impact. First, the international crisis is contributing to the development of a wider and qualitatively new role for the government in economic affairs; there is an intensification of direct government involvement in foreign economic affairs and a new emphasis placed on supply or cost factors in domestic fiscal policy. Second, and in part related to this changing role of government, the international crisis is effecting an intensification of class struggle within the United States.

This paper is intended as part of a Marxist interpretation of the current economic crisis. In this context, we would like to call attention to four main themes that underlie our analysis: 1) the fundamental unity of capitalism as a world system, (2) the class nature of the capitalist state, (3) the development of crises out of the basic contradictions of the capitalist production process, and (4) the interaction between changes in economic relations and class struggle.

## I. The Crisis and Shifts in International Business Activity

Ironically, the increased instability of international affairs has been accompanied by a heightened intensity in *U.S.* business's foreign activity. During the 1970's, the rate of growth of direct foreign investment has maintained its high level of the 1960's. The growth of both imports and exports has substantially outstripped the increase in *GNP*.

There are two explanations for this rise in international activity in time of crisis. First, the crisis internationally is bound up intimately with the domestic economic dislocations which have driven down profit rates within the United States significantly over the past decade. Consequently, the search for profits abroad through both trade and investment has assumed a new urgency. Second, the changed international environment has been characterized by the increased competitiveness of foreign manufactured goods in *U.S.* markets, as well as by a sharp rise in the prices of petroleum products. Both have contributed to an expansion in imports; and, in an effort to alleviate the consequent *U.S.* trade deficits while also maintaining the value of the dollar, the government has promoted exports with unparalleled vigor.

It is in examining some of the particu-

* Graduate student, University of Massachusetts–Amherst, and Visiting Associate Professor of Economics, University of Massachusetts–Boston, respectively.

lars of these aggregate responses to the crisis that the most interesting implications come to light. We shall look, in turn, at import shifts, export shifts, and changes in the pattern of direct foreign investment. Each of these is associated with a qualitative increase in the government's direct involvement in economic affairs.

### *A. Import Shifts*

Over the past several years, accompanying the dramatic rise in total imports, there have been two sorts of shifts in the composition of imports. The first of these is the increasing importance of manufactures in the import bundle. In the 1960-64 period, 46 percent of merchandise imports to the United States were manufactured goods; in the 1965–69 period, the figure rose to 59 percent; and for 1970–74, it stood at 63 percent.[1]

This first change in import activity provides an accounting of the process by which the other capitalist nations have overcome the post-World War II industrial supremacy of *U.S.* business. While the relative advance of those nations had been on the economic horizon for a number of years, slow productivity growth in the United States during the late 1960's and the associated inflation led to a marked increase in the pace of that advance.

The second major shift in the composition of *U.S.* imports has been the higher share represented by petroleum and petroleum products. As a share of total imports, crude materials and fuels—mostly oil—fell from 28 percent in the 1960-64 period to 20 percent in the 1965–69 period, and then dropped still further to 16 percent for 1970–72. But for 1973–74, crude materials and fuels rebounded to 26 percent of total imports.

The higher share of crude materials and fuels has been brought about largely by the jump in the price of petroleum, rather than by any increase in the quantity of petroleum imports. The price increase has, of course, been the result of a significant restructuring of world power relations. With its developing defeat in Vietnam and the concomitant weakening of the *U.S.* economy, the government was in no position in 1973 either to discipline the oil-producing nations or to bring about a common response on the part of the major oil-consuming countries.

The combination of the long-term increase in manufactured imports and the recent sharp rise in the value of petroleum imports has presented the United States with severe difficulties.[2] At the very least, these changes have contributed to the destruction of *U.S.* hegemony in world capitalist affairs. The dollar basis of the international monetary system, for example, has been torn apart by the relative growth of foreign business strength. This has meant, and will likely continue to mean, substantial losses for *U.S.* businesses. However, the *U.S.* government has taken strong action in an effort to dampen the impact of rising imports. (In the absence of such government involvement, there would be an even further decline in the value of the dollar and an even greater transfer of real income from the United States to foreign countries.)

At least three sorts of actions have been taken by the government. The first and most visible of these has been the recent attempt to formulate a comprehensive national energy program. While we are not

[1] In 1973 and 1974, the huge increase in the cost of oil imports dampened this trend in import composition. If we look only at the three years 1970–1972, manufactures accounted for 67 percent of total imports.

[2] To a degree, the international petroleum situation has helped to offset the relative *U.S.* decline with regard to trade in manufactured goods because the other capitalist nations have been hurt more seriously by the rise in oil prices. Nonetheless, while the "energy crisis" has meant a relative setback for European and Japanese business, it has not halted the continuing challenge to their *U.S.* competitors.

concerned here with the details of the conflicts surrounding this attempt, we do wish to note that, regardless of what particular policy emerges, that policy and the process of its formulation will have involved a high degree of government participation in the energy sector of the economy.

The second consequence of the rising importance of oil imports has been an effort on the part of government and the oil industry to diversify sources of imported oil. Nigeria has quickly become one of the largest suppliers of imported petroleum to the United States. Indonesia and Angola (the latter still under Portuguese control in this period) also became important suppliers in the early 1970's. These developments could hardly have taken place without a concerted effort by Washington to promote "friendly" regimes in these nations. While the involvement of the *U.S.* government in helping to obtain and protect foreign sources of supply is hardly new, the oil crisis has forced an intensification of government concern. As a tactic in combatting petroleum supply difficulties, though, increased government involvement should be seen as merely changing the nature of the problem: the potential political conflict arising from the situation in Angola, for example, is hardly less serious than the oil crisis itself.

### *B. Export Shifts*

The third area of government response to the problem of rising imports has been a vigorous policy of export promotion. The trade deficit which first appeared in 1971 came as no surprise since it merely represented an extension of trends established during the preceding several years. Accordingly, it is also not surprising that plans for responding to the deficit were already in the making.[3] Washington foresaw the relative decline in the export potential of the *U.S.* manufacturing sector and, as part of the remedy, laid out a program for the expansion of *U.S.* agricultural exports. The two central aspects of the program were:

(1) the negotiation of special arrangements with foreign governments. This aspect of the program has been carried out by means of such expedients as the Soviet grain deals and pressure on European governments to adjust trade regulations in favor of *U.S.* agriculture.
(2) the structuring of domestic prices to favor agricultural production. This aspect of the program was set in motion through the wage-price controls which took effect in August 1971. Raw agricultural products and exports were explicitly exempted from controls, encouraging the growth of farm output and its sales in foreign markets.

The program worked. The value of agricultural exports tripled from 1970 to 1974. Between 1966 and 1971, agricultural goods accounted for 18 percent of *U.S.* exports; for the period 1973, 1974, and the first half of 1975, the figure has been 23 percent.

The second part of the export-promotion strategy consisted of a policy of vigorous sales of armaments to foreign governments. *U.S.* arms sales abroad climbed from \$1.7 billion in fiscal 1971 to \$9 billion in fiscal 1975, or from 4 to 10 percent of total exports.

The intimate relation of the *U.S.* government with arms producers in the promotion of military sales abroad hardly needs documentation here. The govern-

[3] This foresight took the form of a report, *United States International Policy in an Interdependent World*, prepared by a special Presidential Commission and published in 1971. For a discussion of the report and of the "planned" nature of the rise in food prices, see Boddy and Crotty.

ment both finances weapons research and development costs and facilitates marketing as part of the normal process of foreign policy.

We have a situation, then, in which the two main thrusts of *U.S.* export expansion—agriculture and arms—have been implemented with the planning, sponsorship, and active intervention of the government. Though government involvement in foreign trade is nothing new, the degree and importance of that involvement in response to the current crisis suggest a qualitative change in the relationship of government to the economy.

### *C. Investment Shifts*

As with trade, government involvement with the process of foreign investment is not new. Yet here, too, current developments point to increasing intensification of that involvement. In particular, there seems to be a shift toward more manufacturing investment in certain underdeveloped countries. The evidence on this point is still fragmentary and can only be taken as suggestive. The share of total direct foreign investment classified as manufacturing in "developing" countries has crept up from 7 percent in 1966 to 8 percent in 1974. However, because the dollar has declined in value relative to the currencies of most advanced countries but has remained more stable relative to those of nations in the "developing" category, the real shift toward the "developing" nations has surely been greater than these figures indicate. In certain countries, the recent increase in manufacturing investment has been very large. Total *U.S.* manufacturing investment in Brazil, for example, jumped from $2.03 billion in 1973 to $2.50 billion in 1974; and in Mexico, from $1.80 billion to $2.15 billion.

This apparent new growth of manufacturing investment in certain underdeveloped countries is a direct consequence, on the one hand, of rising unit labor costs in the advanced capitalist nations and, on the other hand, of the development of large and "well-controlled" work forces in certain poor countries. South Korea and Brazil are among the oft-cited examples, the latter recently described by *Business Week* as a country in which "More than a decade of authoritarian military rule has created a climate of political confidence and virtually eliminated labor problems. . . ."

The trouble with military regimes, though, is that even the most stable among them require continuous injections of military support from Washington. And often they are not so stable. Accordingly, as with the attempt to diversify sources of oil supplies, current developments in the foreign investment field are tying the *U.S.* government ever more firmly to policies of manipulation and intervention throughout the world.

## II. The Crisis and the Domestic Role of the State

The current international crisis has not only led to intensified government involvement in the international activity of *U.S.* business. It has also brought about a changing government role in domestic affairs.

The changing role of the government is most usefully interpreted in the context of the central contradiction surrounding the value creation and realization process in capitalist society. In its efforts to maximize surplus-value, business is hemmed in on two sides. On the one hand, failure to suppress wage costs directly and immediately cuts into surplus-value. On the other hand, success in suppressing wages undermines the expansion of effective demand, the basis for the realization of surplus-value.

During the long expansion of the 1950's and 1960's, the manifestation of this con-

tradiction was averted by two sets of circumstances. First, the government contributed heavily to the maintenance of high levels of effective demand. Second, low fuel and food costs, both made possible by *U.S.* hegemony in international affairs, allowed business wage costs to remain low. Moreover, government spending to expand demand combined with limited foreign competition to make possible the simultaneous growth of profits and wages.

But by the late 1960's, the expansion of government spending had contributed, in two ways, to the undermining of profit rates (and rates of surplus-value). First, unemployment had dropped to relatively low levels, increasing the bargaining power of labor and making for a shift in factor shares away from capital and toward labor. Second, expenditures on welfare, unemployment compensation, and other social services had exerted upward pressure on wage rates. As the crisis in international affairs has simultaneously increased food and fuel prices and increased foreign competition, the problems created by the expansionary policies of the 1950's and 1960's have become more severe.

A partial remedy is being sought in limiting the size of the government deficit. However, given the present depressed state of the economy and the fear of another downturn in the near future, the potential for reduction of the deficit is severely limited. It seems likely therefore that greatest attention will be given to developing a fiscal strategy that will maintain aggregate demand while placing a minimum of upward pressure on wages.

On the spending side, this strategy calls first for a reduction in government spending on social service programs. Cases in point include the administration's recent efforts to cut back food stamps, government job creation, day care and child nutrition programs, and other aspects of health and education support. Federal action is being coupled with state and local efforts around the country to cut back on welfare payments and aid to the unemployed. Even the nature, if not the level, of military expenditure is likely to be affected, with more emphasis on research and development, for example. The tax structure implicit in the government's developing strategy hardly needs to be stated: there will certainly be an effort to shift the burden away from capital toward labor.

Changes in policy orientation on the scale of those currently being undertaken require thorough ideological support. The barrage of propaganda concerning a "capital shortage" is a central element in forging the necessary ideological climate. To be sure, the "capital shortage" is not simply fictitious; the international and domestic crises have meant real, structural changes in the economy affecting production costs. The issue, however, is who will pay the new costs. Were the "capital shortage" propaganda to work, labor would accept the shifts in government fiscal policy and would acquiesce at the bargaining table.

The "capital shortage" ideology provides a lever for dealing with a number of problems. For example, it seems to be working well as a guiding principle for handling the New York City crisis as well as other local fiscal difficulties. In New York, business and government have united (in spite of sectarian tactical disputes) around a program which would accomplish at the local level, and in an exaggerated manner, what the federal budget would accomplish at the national level. Gains won by labor in contract struggles and in political battles would be sharply cut back by any of the proposed "assistance" programs.

In sum, the "capital shortage" rhetoric is a major part of the ideological cam-

paign at the center of an intense political and economic offensive currently being waged by business interests in response to the economic crisis.

## III. Conclusions

The business offensive is, however, likely to generate a response. There is little reason to expect labor and various social action movements to abandon their militance simply because business is making threats. The bureaucratic nature of organized labor may serve to dampen militance. But the power of union bureaucracies has been based on an ability to deliver the goods, and now that ability has been severely undermined.

Furthermore, labor's response to the current business offensive may take on some qualitatively new aspects. The depoliticization of working class struggles in the United States has depended on more than a steadily rising standard of living. Another major factor has been the ability of the government to maintain an illusion that it plays a class-neutral role in economic affairs. The crisis itself has interrupted the advance of living standards. And, as we have argued in this essay with emphasis on international affairs, the government is being forced by the crisis to take a more direct and active role in the economy. It is thus becoming less and less tenable for the government to retain its class-neutral image. Accordingly, we think it likely that as working people develop a response to the current crisis, their actions will become increasingly politicized and class conscious.

## REFERENCES

R. Boddy and J. Crotty, "Food Prices: Planned Crisis in Defense of Empire," *Socialist Revolution,* Apr. 1975, *23,* 101–09.

# THE LAST GREAT DEPRESSION AND THE PRESENT ONE

# Lessons for the Present from the Great Depression

*By* PETER TEMIN*

The economic contraction that started in 1929 was the worst in history. Historians have compared it with the downturns of the 1840's and the 1890's, but the comparisons serve only to show the severity of the later movement. In the nineteenth-century depressions, there were banking panics, deflation and bankruptcy in various proportions. But there is no parallel to the massive underutilization of economic resources in the 1930's.

Given the magnitude and importance of this event, it is surprising how little we know about its causes. The reactions of people to the Depression, the policies undertaken during the Depression, and the effects of the Depression have all been the objects of extensive study. But the economic collapse itself has suffered a form of intellectual neglect. Too long ago to be part of the study of the current economy, too recent to be included in most courses in economic history, too complex to be explained simply, economists have—with a few prominent exceptions—left the study of the Depression to others. The inevitable result has been a neglect of the economic aspects of the Depression. While economists have advanced a variety of hypotheses to explain why depressions can take place, little attention has been given to the explicit application of these competing theories to the biggest depression in history.

One reason, often and correctly given for the magnitude of the Great Depression, is the absence of a concerted expansionary macroeconomic policy between 1929 and 1933. The monetary and fiscal policies that we now think could have been effective in moderating or eliminating the contraction were not used to any perceptible extent.[1]

But the question of what macroeconomic policy would have worked is only one question that can be asked about the Great Depression. It is of considerable interest also to know what happened to make such a corrective policy desirable.[2] What happened in the years around 1929 that (in the absence of offsetting govern-

* The research on which this paper is based was supported by the National Science Foundation. The paper has benefited from the comments of Rudiger Dornbusch, Stanley Fischer, and Frederic Mishkin. All are to be thanked. None are to be implicated in the views expressed. The views expressed here are the author's and do not reflect those of the Dept. of Economics or the Massachusetts Institute of Technology.

[1] For the failure of monetary policy to be used, see Friedman and Schwartz, 1963a. For the failure of fiscal policy, see E. Cary Brown, 1956. M. R. Norman, 1969, argues that a successful counter-cyclical fiscal policy would have had to exceed greatly the scale of anything even contemplated in the 1930's. Of course, the negative argument that macroeconomic policies were not used is hardly the same as the positive argument that these policies would in fact have been effective. This assertion is difficult, possibly even impossible, to prove, and the debate about it revolves on questions about the structure of the economy rather than the occurrence of specific historical events.

[2] This is the question I have addressed in Temin, 1976.

ment policies) led to the historically unique events of the 1930's? While this question is different from the question of what policies are appropriate for curing depressions, it is not unrelated, as the work of Milton Friedman and Anna Schwartz shows.

Friedman and Schwartz's classic *Monetary History of the United States* forms the base of an argument that movements of income have historically been caused by changes in the stock of money, which in turn are caused by a variety of exogenous factors.[3] The *Monetary History* documents the last step of this argument, from which the other steps are derived.

Upon examination, however, the *Monetary History* turns out to be an account of the supply of money. Friedman and Schwartz never acknowledge the existence of the identification problem implicit in any attempt to partition changes in a quantity into changes in the supply and changes in the demand. Instead, they discuss the supply of money and ignore the demand—a procedure that can only be justified by the assumption that changes in the stock of money were attributable almost entirely to shifts of the supply curve. This proposition, which the *Monetary History* appears designed to prove, is instead an assumption on which the analysis of the book rests.[4]

For instance, the *Monetary History* asserts that the stock of money fell in the early 1930's because of an autonomous fall in the supply of money. The fall in the supply was caused in the first instance by a decline in Federal Reserve credit outstanding in 1930, but far more importantly by the effects of the banking failures that started in late 1930. The fall in the supply of money induced a movement along the (stable) demand curve for money—that is, a decline in income and interest rates—to equilibrate the money market.[5] The fall in income therefore was a result of the autonomous factors decreasing the supply of money.

Because of the incompleteness of the book's argument, however, there is nothing in the narrative of the *Monetary History* to refute the following completely different story: Income and production fell after 1929 for nonmonetary reasons. Since the demand for money is a function of income, the quantity of money demanded fell also. The money market was equilibrated by the fall in the stock of money following the bank panics, but these panics did not in any way cause the decline in income.[6]

It is not easy to resolve the implicit identification problem and choose between these two stories, and I have tried a variety of approaches. Space precludes anything more than a brief mention of my conclusions here as they are fully reported elsewhere (see Temin, 1976). I found that it is harder to discriminate between these two stories than one might imagine from the intensity of the debate among macroeconomists, but that there is no evidence of any effective deflationary pressure from the banking system between the stock-market crash in October 1929, and the British abandonment of the gold standard in September 1931.

As mentioned at the start of this discussion, this disagreement among histori-

[3] Friedman and Schwartz, 1963a. The argument is implicit in the *Monetary History*, but made explicit in Friedman and Schwartz, 1963b, and Friedman and David Meiselman, 1963.

[4] This interpretation of the *Monetary History* is extensively documented in Temin, 1976, Chapter II.

[5] Friedman and Schwartz's argument assumes that the bank failures had a much stronger effect on the demand for deposits than on the demand for money. For a first approximation, the latter may be ignored.

[6] In Friedman and Schwartz's story, income would not have declined nearly as much as it did in the absence of the banking panics. In this alternative story, income would not have been much higher in the absence of the panics.

ans is of more than passing interest. Our inability to settle on a single story of the Great Depression, or even to devise powerful tests of alternative stories, suggests that a certain diffidence with respect to current policy prescriptions is in order. It also makes it hard to generalize from the experience of the 1930's because it is not clear which aspects of that time were critical in the economic decline. Nevertheless, a stab at comparison is worth making. Is there anything in current macroeconomic data that suggests that we are in for a period of substantial and sustained underutilization of economic resources on a scale approaching the experience of the 1930's?

Despite current optimism on this score, there are a number of similarities between the recent decline and the beginning of the Great Depression. Some of them are shown in Table 1, where annual data is used in preference to quarterly data to make the recent observations consistent with the available annual data from the 1930's. Annual rates of change of several macroeconomic variables are shown for the first two years of the Great Depression, for last year, and this year. The data for 1975 have been estimated from data on the first three quarters, but they should be accurate to the degree needed for this discussion.

The *GNP* deflator is listed in the first row of Table 1 because recent price history is so different from the price history of the late 1920's and early 1930's. It seems absurd to discuss, say, the nominal quantity of money without paying attention to the rate of inflation, but there are also objections (as we shall see shortly) to dealing solely with real balances. I therefore have shown the changes in a price index separately to emphasize the price movements removed in the calculation of real magnitudes.

The second and third rows of Table 1 show the growth rate of the real stock of money under two definitions. In nominal terms, the path of the money stock in the last two years was completely different than in 1929–31. In real terms, however, there is more similarity. The real money stock remained more or less constant in 1929–30 and rose in 1930–31; it fell slightly in 1973–74 and 1974–75. In neither period did the real money stock behave in an expansionary manner, and it fell more in the last two years than it did in the early 1930's.

TABLE 1—PERCENTAGE CHANGES IN SOME MACROECONOMIC VARIABLES, 1929,31 AND 1973,75 (Percent)

| | 1929–30 | 1930–31 | 1973–74 | 1974–75 |
|---|---|---|---|---|
| *GNP* deflator | −3 | −9 | +10 | +9 |
| Real $M^1$ | −1 | +3 | −4 | −4 |
| Real $M^2$ | +1 | +3 | −1 | −1 |
| Real stock prices | −16 | −26 | −36 | −7 |
| Real residential construction, 1958 prices | −39 | −19 | −27 | −22 |
| Real consumption expenditures, 1958 prices | −7 | −4 | −2 | +1 |
| Real *GNP*, 1958 prices | −10 | −8 | −2 | −3 |

*Sources:* Council of Economic Advisors, *Annual Reports*, 1975, 1971; Board of Governors of the Federal Reserve System, *Banking and Monetary Statistics* (Washington, 1943); Standard and Poor, *Trade and Securities Statistics*, 1974 and 1975; and *Survey of Current Business*, recent issues. $M^1$, $M^2$, and stock prices were deflated by the *GNP* deflator. The last column shows estimates made informally from data on the first three quarters of 1975.

What should we conclude from this comparison? If we believe that the demand for money is a stable function around which there is very little variation, that the interest-elasticity of the demand for money is low, and that the direction of causation runs from the stock of money to the arguments on the right-

hand side of the familiar equation, we must look for a fall in real *GNP*.[7]

This approach ignores interest rates and other financial markets; only an extreme monetarist would stop here. One financial market whose movements have been cited repeatedly as a cause of the Great Depression is the stock exchange. As can be seen in Table 1, prices on the New York Stock Exchange declined more from 1973 to 1974 than they did from 1929 to 1930, and the difference is quite large in real terms. And even though stock prices fell less in real terms this year than in 1931, the fall over a two-year period was almost exactly the same for 1929–31 and 1973–75. If the stock-market crash was a powerful deflationary force in 1930, there would seem to be cause for alarm now from this quarter also.

Turning to nonfinancial markets, the decline in housing in the 1930's often is cited as a primary cause of the Great Depression. And, as everyone knows, the construction industry is a serious casualty of the current financial picture. As Table 1 shows, the proportional decline in the real value of housing investment was only slightly smaller in 1973–74 than in 1929–30 and slightly larger in 1974–75 than in 1930–31. The data in Table 1 are only the tip of the iceberg; the history of construction is too complex to be summarized this simply. But a more detailed history would not refute the assertion that *if* a prosperous construction industry is needed for the health of the economy, as most Keynesian descriptions of the Depression seem to assume, then the recent decline in housing is extremely serious.

[7] This argument typically is made in nominal terms by monetarists, and the inflation rate is then subtracted from the growth of nominal income to get the growth rate of real income. If the income elasticity of the demand for money is one, that is, if velocity is constant as is often assumed, the two procedures are the same.

(See Ben Bolch and John Pilgrim.)

So much for similarities. There are also several striking discrepancies between the recent record and the experience of the 1930's differences that we hope are enough to justify the current optimism. The instability of foreign exchange markets, the difficulties attendant upon large shifts in international lending, and the problems of countries experiencing sharp changes in the terms of trade, all figure prominently in many stories of the 1930's. All of these problems are present today, but the structure in which they are taking place does not seem as vulnerable as the inter-war one. In particular, the era of floating exchange rates does not seem to be as vulnerable as the gold-exchange standard was.

In addition, the fall in real consumption expenditures was far smaller in 1973–74 than in 1929–30. The dramatic decline in consumer spending in 1930 was an important—possibly even critical—part of the story of the Great Depression. In my view, an autonomous fall in consumer spending in 1930 was a major factor blocking the recovery that contemporary observers expected until quite late in that year (see Temin 1976, Chapter III). If this view is accepted, then the current relative steadiness of consumption expenditures is quite encouraging.

This raises an obvious question: Why didn't consumer spending fall as much relative to the predictions of a standard set of expenditure functions last year as it did in 1930? A tentative answer is suggested by some recent work of Frederic Mishkin (1975a and b). Consumer expenditures, Mishkin argues, are related to the balance-sheet position of households, in which the stock of money figures only marginally. It has by now become a commonplace that household spending plans are in part a function of the household's net wealth. In addition, Mishkin argues

that the costs of selling illiquid assets—or in the extreme case of going bankrupt—induce household spending to be a function of the household's leverage as well. At the same level of net wealth, consumption expenditures will be inversely related to the volume of debt.

People were increasing their indebtedness throughout the 1920's and they borrowed heavily in 1929. When the stock-market crash came, the net wealth of households was reduced and their leverage was *increased*. (The decline in the value of their assets both reduced their net wealth and raised the ratio of their debts to their assets.) The combination of these two movements caused consumption expenditures to slump dramatically in 1930.

This effect was not present to the same degree in 1974 for two reasons. There was no accumulation of debt in 1973 comparable to the build-up in 1929. And the inflation of 1974 reduced the real value of debt while the deflation of 1930 increased it. Consequently, while the effect of the decrease in consumer liquidity was apparent in 1974, it did not have the dramatic macroeconomic implications that it did in 1930.

Finally (in our list of optimistic factors), there does not seem to be a substantial risk of repeating the sequence of bank failures that characterized the early 1930's. The *FDIC* was able to contain the failures of the United States National Bank of San Diego in 1973 and the Franklin National Bank of New York in 1974 in ways that aborted any rise in destabilizing speculation against banks. There does remain—at the time of writing—the risk that a default by New York City might weaken banks holding the city's paper and lead to a banking panic, but this risk appears small. The Federal government probably will not let the city default, and investors probably will be happy with a negotiated write-down of assets in place of a court-directed write-down. Even if the city does default, this possibility has been extensively discounted. An event so fully anticipated should not initiate a panic, although it may have less dramatic contractionary effects by raising the interest rates at which many potential purchasers of real capital can borrow.

Our view of the Great Depression is hardly complete. In a serious discipline, it must be a cause of concern that one of the major events in our economic history is still so poorly understood. Nevertheless, our knowledge of the Great Depression is still sufficient to give us some assurance that the current position is not nearly as serious as the economic condition in 1930. It would be nice to say that this was the result of more enlightened macroeconomic policies by our current leaders, but the story just sketched does not support such a conclusion. We are the beneficiaries of some changes in the structure of the economy and luck rather than the fruits of expert leadership.

## REFERENCES

**B. Bolch and J. Pilgrim**, "A Reappraisal of Some Factors Associated with Fluctuations in the United States in the Interwar Period, *Southern Econ. J.*, Jan. 1973, *39*, 327–44.

**E. C. Brown**, "Fiscal Policies in the Thirties: A Reappraisal," *Amer. Econ. Rev.*, Dec. 1956, *46*, 857–79.

**M. Friedman and D. Meiselman**, "The Relative Stability of Monetary Velocity and the Investment Multiplier in the United States," in *Stabilization Policies*, New Jersey 1963.

——— **and A. J. Schwartz**, *A Monetary History of the United States 1867–1960*, Princeton 1963a.

——— **and A. J. Schwartz**, "Money and Business Cycles," *Rev. Econ. Statist.*, Feb. 1963b, *45*, 32–78.

**F. S. Mishkin**, "Illiquidity, Consumer Durable Expenditure and Monetary Policy," *Amer. Econ. Rev.*, 1975a (forthcoming).

———, "The Household Balance Sheet and the Great Depression," 1975b (unpublished).

M. R. Norman, *The Great Depression and What Might Have Been: An Econometric Model Simulation Study,* unpublished Ph.D. dissertation, Univ. of Pennsylvania 1969.

P. Temin, *Did Monetary Forces Cause the Great Depression?* New York 1976.

# The Demand for Money from the Great Depression to the Present

*By* ARTHUR E. GANDOLFI AND JAMES R. LOTHIAN*

Myths and legends about the Great Depression have dominated the public's perception of the business cycle. They have shaped government policy and, until recently, they have even held powerful sway among economists. In the immediate aftermath of the Great Depression, many economists came to question the fundamental concept of economic equilibrium. Velocity was thought to be highly unstable—as in the case of the "liquidity trap"—so that the quantity of money supplied was consistent with any level of nominal income. In the real sector, there was the frightening specter of underemployment equilibrium. Following a shock to the economy which reduced output and depressed prices, there were reputedly no forces at work that would achieve a return to full employment. What is remarkable about these ideas is that they gained such widespread acceptance in the economics profession despite the absence of any systematic supportive evidence.

As empirical evidence on the economy accumulated, more and more economists turned away from such views. Milton Friedman and Anna Schwartz's explanation of the monetary nature of the Great Depression provided the major stimulus for this reevaluation. Their conclusions were buttressed by the results of other statistical studies which have dealt directly with the 1930's, such as those by Martin Bronfenbrenner and Thomas Mayer, by Allan Meltzer, by Clark Warburton and most recently by Arthur Gandolfi. Their arguments have been given support by the numerous studies of money demand that have demonstrated its essential stability under diverse institutional and economic conditions.

The current recession has led to a renewed concern over economic instability. This concern is evidenced by the resurrection of a whole list of proposals reminiscent of the 1930's—running the gamut from *WPA*-type employment projects to national economic planning. The Great Depression again is pointed to as the premier example of economic instability; the current recession as its second coming. One reason for the persistence of the controversy surrounding the Great Depression lies in the difficulty in testing competing explanations of the behavior of the economy during short historical episodes. The number of truly independent time-series observations is very limited and the degree of collinearity among possible explanatory variables is too high.

Our paper focuses on one of the controversies noted above—the behavior of the demand for money during business contractions. This issue is of particular importance since a stable demand function for money is an essential requirement for a monetary explanation of the cycle. To get around data limitations, which result from the use of time series alone, we use yearly cross-state data for the period 1929 through 1968. This approach increases

* First National City Bank. We are grateful to John Maher for his extremely competent assistance in the preparation and analysis of the data. We also want to thank Thomas K. Barneby, Phillip Cagan, Stanley Diller and Anna J. Schwartz for their comments on an earlier draft and Deborah Wenninger for her programming assistance.

our degrees of freedom and reduces problems of collinearity. It also provides the possibility of reconciling the pre- and post-World War II movements of velocity which have made it difficult to say anything definitive about velocity and the demand for money over short time periods like the early 1930's.[1]

## I. The Model and the Data

As a model of the demand for money we assume a fairly standard relationship of the form:

$$ln\,(M/PN) = \beta_0 + \beta_1\, ln\, y_p + \beta_2\, ln\,[(1 + RB)/(1 + RM)],$$

where $(M/PN)$ is per capita real total commercial bank deposits, $y_p$ is permanent per capita real personal income and where the bracketed term is the ratio of the value in the next period of a dollar invested in long-term bonds to the value in the next period of a dollar held as deposits—a measure of the opportunity cost of holding money.

The data we use cover the continental United States and the District of Columbia for the forty years from 1929 through 1968. Real per capita money holdings are estimated by deflating total commercial bank deposits for each state by the national consumer price index and by total state population. In calculating the state permanent income series, we used an exponentially declining weighted average of state per capita personal income with an initial weight of .33 and an allowance for separate state trends in income. The rate of interest paid on deposits was calculated by dividing the total interest paid on deposits, net of services charges on demand deposits, by a weighted average of June and adjacent December deposit figures.

## II. Empirical Results

We obtained estimates of the parameters of our money demand function from a wide variety of regressions—individual yearly cross-state regressions and pooled regressions, for the full 40-year period and for different subperiods.[2] We find that on the whole the demand for money has been stable over these 40 years. The standard errors of our regressions show little variation whether we go from period to period or whether we aggregate temporally. And the income and interest rate coefficients obtained from regressions run over the five-year and ten-year subperiods are remarkably stable. These regressions are presented in Table 1.

The income coefficients from the five-year regressions range from 1.20 to 1.42 and, for the 10-year regressions, from only 1.29 to 1.35.[3] Interest rate coefficients indicate an inelastic response of money demand to interest rates in all instances. In the five-year regressions, however, the effect of interest rates on the demand for money is considerably less stable than that of income. The coefficients themselves go from –4.74 to –15.70; the elasticities (estimated at the mean) from –.13 to –.53. But in the 10-year regressions this

[1] Prior to the war, velocity in the United States—particularly that of $M^2$—trended downward, suggesting to many observers a secular decline. In contrast, the velocity of $M^1$ increased more or less steadily since 1946, while the velocity of $M^2$ increased from 1946 until the early 1960's, remaining relatively constant thereafter. Not surprisingly, the bulk of the income elasticities estimated in demand for money studies with postwar data have been markedly lower (.5 to .7) than the longer term estimates (1.0 to considerably above) covering both the prewar and postwar periods.

[2] All regressions which include the years 1929 through 1936 contain current and lagged failure rate variables. For justification of the variables see Gandolfi's earlier paper.

[3] Our finding that the income elasticities estimated cross-sectionally are both stable and greater than unity casts considerable doubt on the hypothesis that the postwar rise has been caused by a decrease in the income elasticity of demand. It implies that the low estimates of income elasticities others have found using postwar data are the results of specification bias rather than any systematic alteration in income elasticity.

Table 1—Regressions of $ln\ (M/PN)=\beta_0+\beta_1\ ln\ y_p+\beta_2\ ln\ [(1+RB)/(1+RM)]$[a]

| Period | $\beta_0$ | $\beta_1$ | $\beta_2$ | $SE$ | $\bar{R}^2$ | $\epsilon_r$ | Period | $\beta_0$ | $\beta_1$ | $\beta_2$ | $SE$ | $\bar{R}^2$ | $\epsilon_r$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1929–33 | −2.48 | 1.30 | −14.30 | .200 | .89 | −.44 | 1929–48 | −2.69 | 1.29 | −5.62 | .212 | .88 | −.15 |
| | (9.24) | (35.87) | (5.73) | | | | | (19.08) | (75.19) | (3.38) | | | |
| 1934–38 | −2.93 | 1.33 | −4.74 | .213 | .86 | −.13 | 1949–68 | −3.08 | 1.34 | −10.99 | .198 | .70 | −.34 |
| | (10.10) | (36.16) | (1.33) | | | | | (14.78) | (48.01) | (8.95) | | | |
| 1939–43 | −2.91 | 1.34 | −15.26 | .222 | .83 | −.35 | | | | | | | |
| | (9.18) | (34.49) | (2.46) | | | | 1929–68 | −1.94 | 1.20 | −12.97 | .212 | .87 | −.37 |
| 1944–48 | −3.02 | 1.34 | −9.87 | .196 | .75 | −.24 | | (22.57) | (101.25) | (14.37) | | | |
| | (6.32) | (25.74) | (1.33) | | | | | | | | | | |
| 1949–53 | −3.77 | 1.42 | −8.67 | .179 | .77 | −.22 | *Yearly Intercepts* | | | | | | |
| | (9.42) | (28.30) | (2.02) | | | | | | | | | | |
| 1954–58 | −2.69 | 1.29 | −11.64 | .192 | .68 | −.35 | 1929–68 | | 1.31 | −18.80 | .198 | .80 | |
| | (6.07) | (22.61) | (3.53) | | | | | | (81.37) | (11.48) | | | |
| 1959–63 | −1.85 | 1.20 | −15.70 | .202 | .62 | −.53 | 1929–48 | | 1.33 | −17.33 | .202 | .85 | |
| | (3.42) | (18.32) | (4.59) | | | | | | (68.90) | (7.29) | | | |
| 1964–68 | −2.43 | 1.26 | −11.29 | .213 | .55 | −.39 | | | | | | | |
| | (4.11) | (16.96) | (4.48) | | | | 1949–68 | | 1.28 | −20.31 | .194 | .67 | |
| | | | | | | | | | (41.78) | (9.00) | | | |
| 1929–38 | −2.66 | 1.31 | −11.10 | .207 | .88 | −.32 | | | | | | | |
| | (14.01) | (52.75) | (5.58) | | | | | | | | | | |
| 1939–48 | −3.07 | 1.35 | −11.95 | .209 | .85 | −.28 | | | | | | | |
| | (14.33) | (51.88) | (2.68) | | | | | | | | | | |
| 1949–58 | −3.12 | 1.35 | −12.68 | .186 | .73 | −.35 | | | | | | | |
| | (10.92) | (36.01) | (5.63) | | | | | | | | | | |
| 1959–68 | −2.76 | 1.29 | −12.21 | .209 | .62 | −.42 | | | | | | | |
| | (7.15) | (27.83) | (6.05) | | | | | | | | | | |

[a] For 1929–55 bank deposits (as of June 30) came from *All Bank Statistics* and for 1956–68 from the E.4 call reports, both of the Board of Governors of the Federal Reserve System. For 1929–48 interest expenses of banks came from the *Annual Reports* of the Comptroller of the Currency and for 1949–68 from the *Annual Reports* of the Federal Deposit Insurance Corporation. Long-term bond rates came from Standard and Poors, *Trade and Security Statistics;* state population from the *Population Report* of the Bureau of the Census; and both state personal income and the consumer price index from the *Survey of Current Business* of the Bureau of Economic Analysis.

The $\bar{R}^2$ omits the contribution of the yearly intercepts in the last three regressions. The elasticity of the interest rate variable, calculated at the mean, is $\epsilon_r$ and the numbers in parentheses are absolute values of $t$ statistics.

variation decreases markedly, with the coefficients ranging from –11.10 to –12.68 and the elasticities from –.28 to –.42. What the pattern of the interest rate coefficients shows is that the exact opposite of the liquidity trap, reputed to have characterized the 1930's, occurred. Interest elasticities fell as the level of interest rates declined.

To analyze further the stability of our estimates we ran separate yearly cross-state regressions for each of the 40 years. In these regressions, both the income and interest rate coefficients varied more widely than in the five-year pooled regressions. Income coefficients ranged from 1.06 in both 1966 and 1967 to 1.50 in both 1950 and 1951, rising fairly consistently from 1.28 in 1929 to the peak of 1.50 and thereafter declining until they reached low points in 1966–67. The interest rate coefficients were even more widely dispersed than the income, ranging from –30.93 in 1967 to 16.98 in 1946.[4]

[4] We suspect, however, that these estimates overstate the true variability in interest rate coefficients. Since the long-term bond rates are the same for all states, variations in our interest rate variable in any one year will be due totally to differences in the rate of interest paid on money in different states. For years in which these differences are small—such as from the mid-1930's to the early 1950's—it will be difficult to obtain good estimates. Accordingly, the pooled regressions may give a more accurate picture, than the yearly regressions, of the effects of interest rates on money holdings.

The differences over time in our yearly estimates of demand functions are reflected further in results of analyses of covariance. For the period as a whole, and for the subperiods 1929–48 and 1949–68, we rejected the hypothesis of homogeneity of the yearly regressions. The $F$ ratios to test the significance of the temporal differences in both slopes and intercepts were 2.27 and 1.54 for the subperiods and 3.02 for the whole period. None of these is of overwhelming magnitude, but they are all statistically significant at the .01 level.

What we want to find out is the source of this apparent instability. We wish to discover first whether it is due to a change over time in the relationship between money holdings and the explanatory variables—a change in the income and interest rate coefficients—or to some purely time-related omitted variable which would show up as a change in the intercepts. More important for the purposes of this paper, we want to know whether the apparent instability is essentially a cyclical or a secular phenomenon.

In answer to the first of these questions we found that we could account for the instability in yearly regressions by changes in intercepts rather than changes in the income and interest rate coefficients. For both of the 20-year subperiods and, more importantly, for the period as a whole the differences in yearly slope coefficients became highly insignificant after we allowed for separate yearly intercepts. The $F$ ratio never exceeded .88 in any of these three instances. We also ran a covariance analysis of the differences between the coefficients of interest rates and income in the 1929–48 period versus the 1949–68 period, after allowing for differences in yearly intercepts within each subgroup. The regressions themselves are presented in the bottom half of Table 1. The visual impression of stability is further confirmed by the statistical tests. The differences in slope coefficients between these two subgroups are insignificant at the .05 level—an $F$ ratio of 1.11.

Covariance analyses performed for the 1930's alone indicated homogeneity of slope coefficients for that decade. For the separate five-year periods 1929–33 and 1934–38 and for the two combined, the yearly income and interest rate coefficients showed no significant differences after we allowed for separate yearly intercepts. When we considered slopes and intercepts together we could not reject the hypothesis of homogeneity for the yearly regressions within the subperiod 1934-38 but could reject it—just barely at the .05 level—for the years within the subperiod 1929–33 and within the whole period 1929–38.[5]

To investigate the source of the instability of yearly intercepts we ran several tests. One was based upon regressions with separate intercepts for reference cycles, measured peak to peak. Comparing these regressions with regressions with single intercepts showed that the cyclical intercepts significantly reduced the unexplained variance for both the whole period and for the two 20-year subperiods analyzed separately. Comparing the regressions with cyclical intercepts with the regressions with separate yearly intercepts showed that the variation within cycles remained significant. However, as Table 2 demonstrates, the greater portion of the variations left unexplained by income and interest rates was among cycles and not within. Judged in terms of the mean square due to regression on the dummy variables for cycles relative to the mean

[5] This instability of the intercepts for the first five-year period differs from the results presented by Gandolfi in an earlier article. The difference is due to differences in the treatment of the two purely temporal variables—the interest rate and the price index. Gandolfi's deflator was implicitly permanent rather than the current prices used here. His interest rate variable was $ln\ (1+RM)$, which allows the effects of changes in the numerator of our expression to be picked up by the intercept.

square due to regression on the dummy variables for years (within cycles), it was almost eight times as great for the full period (1.039 versus .136). Judged in terms of the decrease in residual variances it was twice as great (.004 versus .002). Our conclusion is that the instability that we have uncovered is much more a secular than a cyclical phenomenon.

This conclusion is given further credence by an additional test. When we included a dummy variable for the contraction stage of all reference cycles in the regressions with individual cyclical intercepts, in no instance—in neither the full period nor the two 20-year subperiods—

TABLE 2—COMPARISONS OF INTER AND INTRA CYCLICAL VARIATIONS IN THE DEMAND FOR MONEY[a]

| | Residual Variance from Regression With | | | Variations Due to Differences[b] | |
|---|---|---|---|---|---|
| Period | Single Intercept | Cyclical Intercepts | Yearly Intercepts | Among Cycles | Within Cycles |
| 1929–68 | .045 | .041 | .039 | 1.039 | .136 |
| 1929–48 | .045 | .043 | .041 | .662 | .176 |
| 1949–68 | .039 | .038 | .038 | .135 | .082 |

[a] Data sources are given in Table 1.

[b] The variance among cycles is the difference in the sums of squared errors from regressions with single and with cyclical intercepts and the variation within cycles is the difference in the sums of squared errors from regressions with cyclical and with yearly intercepts, both divided by the appropriate degrees of freedom.

was this variable anywhere near significant. Moreover, its coefficient, which measures the deviation in contractions relative to expansions, was never greater than 1.6 percent. This evidence shows that even though there are significant variations in money demand within reference cycles they are not attributable to the cycle per se. They may merely be an extension of the departures we have observed among cycles.

## III. Summary and Conclusions

The stability of the demand for money function is an issue that economists cannot ignore. As long as there exists a stable demand for money function, movements in the supply of money are the crucial determinant of movements in nominal income and prices. Our results, based upon time series of cross-state data, show that money demand has been basically stable both over the whole period we have investigated and in individual years of the decade of the 1930's. And the instability that we did uncover in our estimated demand functions is predominantly secular rather than cyclical. So our results imply that cyclical declines in general—and the Great Depression in particular—neither stem from, nor are aggravated by major structural changes in the demand for money.

Our findings also confirm the efficacy of monetary policy in counteracting substantial cyclical declines. They strongly suggest that, regardless of the initial cause of such a downturn in aggregate demand, changes in money supply can be a useful palliative. The argument by Friedman and Schwartz that the Federal Reserve could have mitigated the severity of the Great Depression seems to us, on the basis of our analysis, to be correct. This does not mean, however, that our evidence justifies continual "fine-tuning." We observe statistically significant short-run departures from our model. Such errors may make countercyclical policy of limited use in anything other than periods of severe economic problems.

Inherent in the Friedman and Schwartz position is the proposition that had the supply of money not been drastically shrunken between 1929 and 1933, the fall in income would have been self-limiting and relatively mild. The alternative explanation, that this decrease in the supply of money was a consequence rather than a

cause of the substantial fall in economic activity, is a logical possibility, but one which we find totally unconvincing. It implies that a fall in the demand for money due, say, to a fall in income will bring about a fall in the supply. There is no obvious reason why this would be the case. To argue so, one would have to demonstrate that the proximate determinants of the supply of money—particularly the deposit-currency and deposit-reserve ratios—are sensitive to the factors influencing the demand. But this argument is inconsistent with the evidence. If demand and supply were interdependent, then in view of substantial differences in conditions of supply over our sample period we would expect to see a good deal of temporal instability in our estimated income and interest rate coefficients. But since that was not the case, we suspect that the degree of interdependence has been small. The evidence accumulated by Phillip Cagan in his analysis of the money supply process leads to the same conclusion. Cagan found that the major movements in the deposit-currency and deposit-reserve ratios during severe reference cycle contractions were the result of panics rather than the falls in economic activity. He concluded further that these panics were themselves episodic and not the products of income declines.

If the influence in the Great Depression ran from income to money rather than the other way around, then it was indeed an historical anomaly marked by fundamental one-time changes in the way the general public and the banking sector formed their preferences. And since we find no evidence of such changes in the overall demand for money we are exceedingly suspicious of the argument that they occurred in these other areas of the economy.[6]

[6] Michael Darby's paper further substantiates this conclusion. He shows that the real sector of the economy had a much greater tendency to move toward full employment in the 1930's than was previously believed.

## REFERENCES


M. Bronfenbrenner and T. Mayer, "Liquidity Functions in the American Economy," *Econometrica*, Oct. 1960, *28*, 810–34.

P. Cagan, *Determinants and Effects of Changes in the Stock of Money, 1875–1960*, New York 1965.

M. Darby, "Three-and-a-Half Million U.S. Employees Have Been Mislaid; Or an Explanation of Unemployment, 1934–1941," *J. Polit. Econ.*, forthcoming.

M. Friedman and A. J. Schwartz, *A Monetary History of the United States*, Princeton 1963.

A. Gandolfi, "Stability of the Demand for Money During the Great Contraction—1929–1933," *J. Polit. Econ.*, Oct. 1974, *84*, 969–83.

A. Meltzer, "The Demand for Money: The Evidence from the Time Series," *J. Polit. Econ.*, June 1963, *71*, 219–46.

C. Warburton, *Depression, Inflation and Monetary Policy*, Baltimore 1966.

Standard and Poors, *Trade and Security Statistics*, New York 1974.

U.S. Board of Governors of the Federal Reserve System, *All Bank Statistics*, 1959.

U.S. Bureau of the Census, *Current Population Reports*, Series p. 25, 1956 and various updates, 1957–68.

U.S. Bureau of Economic Analysis, *Survey of Current Business, Supplement: Personal Income by States*, 1956 and various issues, 1957–68.

U.S. Comptroller of the Currency, *Annual Reports*, 1929–68.

U.S. Federal Deposit Insurance Corporation, *Annual Reports*, 1929–68.

# War-Related Debts and the Great Depression

*By* HEYWOOD FLEISIG*

At the conclusion of World War I, war-related debts were about $12 billion, an amount greater than total *U.S.* private long-term foreign assets, and equivalent to perhaps 15% of *U.S.* national income. Every major western country owed something to someone, but on net most of the war debts were owed to the United States by France, Great Britain and Italy; these four countries in turn, were to receive most of the payments by Germany on the reparations account. The largest single net creditor was the United States, the largest single net debtor was Germany.

The heated and lengthy economic debate about war-related debts, conducted in the context of passionate moral and political disputes, produced two distinct strains of thought: one was the well-known discussion of transfer; the other, less adequately incorporated into the literature, held that the war-related debts critically disrupted the international financial system, possibly started the depression, and probably aggravated it.

No simple, direct line can be drawn, however, from war-related debts to world economic activity: payments on war-related debt were made in the 1920's with no obvious adverse effect on economic activity; payments were cancelled in the 1930's with no obvious beneficial effect on economic activity. Accordingly, most accounts of the world depression center on elements other than war-related debts.

Therefore, this paper begins by raising some analytical issues about the relation between the war-related debts and income, before turning to the role of war-related debts in an explanation of the depression. It then argues that these debts had no important role in starting the world depression, some role in aggravating it, and a central role in the German depression. Although war-related debts did not begin the world depression, the analysis used to reach that conclusion indicates that then-contemporary concern about the economic effects of the debts was probably warranted, that even without a world depression eventual German default was likely, and that such a default probably would have created at least a mild world recession. The paper concludes with a discussion of some implications for the international economy today.

## I. International Debt in the International Economy: Some Conceptual Considerations

First, let us consider the notion of the "ability" to repay a foreign debt. Given the Marshall-Lerner conditions, depreciation by the debtor country will increase its trade balance measured in dollars or gold, the units of account that nominally valued the debts. The debtor cannot, however, increase its trade balance indefinitely. Given a full employment output dependent on techniques and factor quantities, and assuming that physical capital cannot be transferred to satisfy the debt,

* Division of International Finance, Federal Reserve Board, and University of Maryland. I am indebted to Ralph Bryant, Michael Dooley, Edwin Truman, Lance Girton, Dale Henderson, Peter Isard, Anna Schwartz, Jeffrey Shafer, and Mack Walker for several helpful comments and suggestions. The views expressed herein are solely those of the author and do not necessarily represent the views of the Federal Reserve System.

increases in the trade surplus can occur only at the expense of domestic consumption and investment. Minimum acceptable consumption and investment determine the maximum trade surplus and, thereby, the maximum rate at which the debt can be reduced. The debtor's maximum trade surplus will be lower if the creditor takes measures to reduce his counterpart trade deficit, but the maximum will exist even where the creditor does not take such measures.

The maximum rate of repayment need not be the optimum rate of repayment. The optimum rate of repayment is a variant of the now-familiar question of the optimal social savings rate. While the debt-repayment problem has its own interesting features, such as the relation between the terms of trade and the rate of repayment and the imposition of a fixed minimum total social savings equal to the debt imposed, all that need concern us here is that when the optimum rate is less than the scheduled rate, the debtor will optimally roll over the debt with private or public international borrowing.

It is helpful, then, to distinguish three rates of repayment: the maximum rate, the optimum rate, and the rate actually scheduled in the debt. A debtor may be "able to pay" a debt in the sense that the present value of the maximum trade surplus exceeds the present value of the debt. The debtor, nonetheless, must roll over part of the debt if an annual scheduled payment exceeds the maximum annual payment; it will choose to roll over the debt when the annual scheduled payment exceeds the optimum annual payment. The rolling over of war-related debts onto private account, and the accompanying increase in the net U.S. creditor position toward the debtor countries, need represent neither inability nor unwillingness to pay by the debtors, though it was often so criticized during the interwar years.

The roll-over of the original debt can increase the probability of default, particularly if the rolled-over debt is of shorter maturity than the original debt. Consider first a world with only war-related debt whose repayment schedule is fixed and known with certainty. If maximum attainable income and minimum consumption and investment are known with certainty, then the maximum surplus is known with certainty, and default is either certain or impossible. Default is certain if the present value of the debt exceeds the present value of the maximum surplus or, in the absence of private lending, if the maximum surplus in any year falls short of the scheduled payment; otherwise, default is impossible. A probability of default emerges only when these elements have stochastic components; while the repayment schedule may be known with certainty, there will be random components in both income (e.g., crop failures, the effects of foreign recessions, errors in domestic economic policy) and in consumption (e.g., a cold winter).

This probability of default will change when we introduce a private international capital market in which the original public debt can be rolled over. Where default was originally due to an annual scheduled payment greater than an annual maximum surplus, but where the present value of the maximum surplus exceeds the present value of the debt, the private market makes payment possible. However, should maturities shorten as debt is rolled over, the new debt may rise so that the fraction due in any year exceeds the maximum ability to pay. Then the probability of default will also depend on the chance that creditors will recall such a fraction of rolled-over debt in a year or that this debt cannot itself be rolled over. In this way, the probability of default will also depend on stochastic elements in the international loan market.

Whether default mattered in the world depression depends on its consequences as well as its probability. Default in the 1930's had several contractionary effects. First, default was typically preceded by a period during which default was not certain, when the debtor pursued contractionary policies to avoid default. Private debtors reduced consumption and investment; public debtors increased interest rates and taxes and reduced government expenditures. Second, the prospect of default reduced the availability of loans to the debtor, making the measures necessary to avert default more extreme. Third, once the default occurred, there was a period of uncertainty during which *ad hoc* institutional arrangements were made, coupled with stringent expenditure conditions placed on defaulters by lenders making new loans. Since the probability of default is not independent of the rate of interest charged on the loan and there are interest rates at which no rational lender will lend, the conditions we have described may not show up in the interest rate. We may regard the supply curve of loanable funds as backward-bending with respect to the interest rate and shifting with respect to the constituents of default discussed earlier.

In addition to the effects of debt-related flows, we must consider the effects of changes in the real values of the debts themselves on public and private expenditure decisions. Changes in prices and interest rates produced immense changes in the real values of war-related debts. There is only limited indication, though, that national economic policies were set with such international balance sheet considerations in mind: on such grounds some support was given to the appreciation of sterling in 1925 and some opposition offered to depreciation of the mark between 1930 and 1932. There is no evidence that countries that were creditors on war-related debt account expanded income because of increases in the real value of their assets, so we may regard the overall governmental response to increases in the real value of the debts as contractionary.

Private international debt was tied to war-related debt in several ways: private parties held most of the rolled-over public debt, debtor countries sometimes undertook domestic policies to increase the current trade surplus that reduced the ability to pay of private debtors within that country, and policies such as suspension of foreign payments affected the repayment of both private and war-related debts.

Changes in the price level are commonly taken as having no effect on aggregate expenditure through changes in private wealth by assuming equal marginal propensities to consume out of wealth for both debtor and creditor. But analogously treating a change in the probability of default as a change in the distribution of expected wealth would neglect an important asymmetry: while a private debtor will reasonably perceive that expenditure reduction lowers the probability of default, a private creditor will reasonably perceive no connection between his own expenditures and the debtor's default. We may also question whether debtor and creditor view the probability of default equally: massive repatriation by debtors of bonds in default suggests that debtors may have seen the probability of default as lower than did creditors, implying that the expected value of the debt fell more for the creditor than for the debtor.

## II. Explaining the World Depression

A request for an explanation is a request for a causal statement. Since causation is ultimately defined by a model, explaining an event is tantamount to subsuming it under the appropriate model. In building a model of the international economy that

incorporates the features of international debt we have discussed, two difficulties would arise: first, there is a lack of general agreement on what the nondebt features of this model would be; second, assessing the quantitative importance of war-related debts would require the application of estimated parameters to variables at the extremes of the sample range in a model that features some severe discontinuities.

Surmounting these difficulties, we would still find ourselves facing a number of apparently appropriate explanations; we should first agree on some usage concerning the notion of "explanation." Essentially our agreed-upon model would take one set of variables exogenous to the economic system and another set of policy variables and transform them into the endogenous variables whose behavior we seek to explain. The structure that performs this transformation, depending on the period of time for which the model is built, comprises some traits thought immutable, some that are endogenous but change with sufficient slowness to be taken as constant, and some traits that may be changed by government policy.

With such a model we could assess the relative importance of changes in exogenous variables, policy variables, and structure, where we would measure importance by the fraction of lost output for which each element accounts. In doing so we would distinguish explanations of the start of the depression from those treating the length and depth, but we could use this model to untangle those elements of the length and depth that are a longer-term endogenous response to the origin from those elements that revolve around later changes in exogenous variables, policy variables, or structure.

We also require a usage to deal with explanations phrased in terms of counterfactuals, such as "the depression was aggravated by the absence of proper fiscal policy." As a technical economic exercise, our model could readily answer the question of how much less severe the depression would have been had different fiscal policies been pursued. It could equally well tell us how much less severe the recession following the War of 1812 would have been had James Madison pursued countercyclical fiscal policies. But while each is an illuminating technical exercise, the latter is a silly historical question while the former is not. This distinction rests on a historical judgment about the feasibility of the policy action or structural change in question, a problem of induction not substantially more difficult than was the statistical verification of the model.

We lack this splendid model. But discussing it in its near ideal form gives us a rough standard for evaluating the partial and imperfect explanations we are able to posit now. We haven't time here to evaluate all the candidate explanations, eliminating some and buttressing others, so I will tell one plausible story and try to show its relation to the war-related debts.

## III. War-Related Debts and the World Depression

The world depression began with the decline in *U.S.* long-term international lending in 1928 and 1929 and was considerably aggravated by the decline in *U.S.* imports in 1930 and early 1931. Both the decline in lending and in imports were of domestic, not international, origin. War-related debts figured only peripherally during this period, through events in Germany; we will discuss these events separately. At its birth, the world depression was the creature of the United States.

The decline in net *U.S.* long-term international lending was the combined result of an inflow of foreign funds to purchase *U.S.* stocks and a decline in *U.S.* purchases of foreign bonds, in turn due to the *U.S.* stock market boom and accompanying

restrictive Federal Reserve policies. The decline in *U.S.* long-term lending led to mild recessions in Europe directly through increased interest rates and indirectly through reduced imports from Europe by third countries who also borrowed in the *U.S.* market; both in Europe and among third countries restrictive policy actions were also undertaken to maintain gold-convertibility in the face of declining *U.S.* lending.

The decline in *U.S.* imports in 1930 and 1931 followed directly from the decline in *U.S.* income. This decline in income is not well-explained. Two leading candidate explanations are the wealth effect of the stock market crash and a decline in residential building; these, in turn, require explanation, but it suffices for us to note that promising trails do not lead to foreign or international factors. The decline in *U.S.* imports led to reductions in incomes in Europe and among third countries through channels similar to those described for the drop in lending. The decline in aggregate demand was now sufficient, however, to sharply reduce the absolute and relative prices of primary commodities and to place in question the repayment of debt by primary producing countries and regions. Of particular interest to us are the agrarian regions of the United States, Germany and Austria.

Payments on war-related debts were small relative to these changes and it is unlikely that their absence would have significantly altered any market or policy response, except that of Germany, from 1928 to early 1931. From early 1931 to 1933 war-related debts cannot be so easily dismissed. It is possible that without the large short-term international debt amassed by German banks in the process of rolling over the war-related debts, the German banking system would have withstood the failure of the Austrian Kreditanstalt, itself largely a function of the fall in agricultural prices. If so, then the thread that runs from the German bank failures to the German suspension of payments and, thence, to the devaluation of sterling and the 1931 *U.S.* bank failures might have been broken earlier. But this is a list of "ifs" quite difficult to assess and it is terminated with a formidable counterfactual explanation based on the question: Why did the Federal Reserve system permit the bank failures?

## IV. War-Related Debts and the German Economy

While war-related debts may not be a major element in explaining the world depression, they are crucial in understanding German economic and political conditions in the 1920's and early 1930's. Whether the German inflation of 1923–24 followed from excessive international transfers that depressed the mark and raised domestic prices or, more plausibly, from government deficit financing of massive relief payments to Ruhr workers who were passively resisting the Franco-Belgian occupation, there is little question that the inflation centered on reparations. It has been suggested that this inflation destroyed the German middle class and paved the way for Hitler. This is an exaggerated story. Hitler attempted the Beer Hall Putsch at the peak of this inflation, attracted a small following, was promptly suppressed by the army, and remained a minor actor in German politics from 1924 to late 1928. Rather, Hitler's rise in power was highly correlated with the rise in German unemployment; his political support came from agrarian and middle class groups substantially damaged by the deflation of 1930 and 1931; his program repeatedly emphasized public works projects for the unemployed and emergency loans and price supports for farmers.

We earlier conjectured that without war-related debts, Germany might have

absorbed the collapse of the Kreditanstalt while pursuing less contractionary policies. It is more certain that war-related debts prevented a more active German response to the depression. The Brüning government faced difficult policy choices from 1929 to 1932. Increased government spending or reduced taxes would have reduced the current account surplus and increased the government deficit. That, in turn, might have indicated an "unwillingness to pay" that would have undermined the attempts by the German government to have reparations reduced on the grounds of the hardship their payment caused Germany. These attempts succeeded at the Lausanne Conference of 1932. An expansionary monetary policy would have had the same effect on the economy and probably on the views of creditors; under the Dawes Plan the Reichsbank had agreed to keep the discount rate above 5 percent. The only remaining route through these difficulties was the devaluation of the mark, a possibility apparently rejected partly because it would increase the burden of the debt measured in German goods but more because it would raise prices. Brüning pursued a brutally contractionary policy in 1930 and 1931.

Apparently perceiving the political dangers of these economic policies, Brüning shifted course in 1932 and began planning massive public works projects. These were delayed in part by the specific objections to them by foreign creditors of Germany and were not instituted until after Von Papen replaced Brüning. Substantial amounts of funds were spent only after Hitler became chancellor; it is, incidentally, such public works expenditures and not spending on armaments and mobilization that account for the bulk of German expansionary fiscal policy from 1932 to 1935.

We have all been taught that reparations had something to do with World War II, but it is useful for economists to recognize that the most plausible channel through which reparations operated was their impact on German unemployment policy. No German government paid reparations enthusiastically, but German voters repeatedly chose the course of the center parties in gradually reducing payments rather than endorsing the right-wing's persistent proposals of repudiation. It would be reckless to explain world history from 1933 to 1939 in terms of German unemployment from 1929 to 1932, but even a small contribution to those events deserves respect and attention.

## V. Summary and Implications for the International Economy Today

In explaining the world depression, we have linked the war-related debts to the absence of more expansionary German fiscal policies and to the chain of bank failures. However, were we to simulate our near ideal model, we would probably find that more expansive German policies would not have changed the world income path substantially, while the bank failure story, itself highly conjectural, loses much force against the historically plausible counterfactual explanation centering on the failure of the Federal Reserve to prevent the U.S. banks from collapsing. The war debts aggravated the depression by creating a structure that produced a larger decline in income from any given shock, but given the largely domestic origins of events in the United States, a world without war debts would still have had a large depression; world depressions, like many other things, do not succeed unless the United State cooperates. All this is consistent with the possibility that the total cost of the depression was so great that even a relatively small contribution might itself have a fearsome consequence.

What is the implication of this for today's economy? History is an uncontrolled experiment and deriving lessons from it through analogy is hazardous. It is better not to proceed by comparing and contrasting war-related debts and oil-related debts. Rather we should regard history as the full-dress acting out of our models, less precise than an econometric representation, but richer in suggesting how our models might be altered to better suit our needs. The analysis discussed earlier, based on the war-related debt episode, suggests that concern over those debts was warranted; that had the depression not intervened, the rolling over of the war-related debt would have made the German economy increasingly sensitive to autonomous disturbances of given size until default eventually occurred. Such a default, with its contractionary preface and aftermath, would probably not have led to anything like the world depression; but we cannot make this argument with great certainty and, moreover, all this is consistent with the absolute loss in output being quite large.

The implication for the oil-related debt is similar. It is a large debt and an optimal pattern of repayment will involve considerable rolling over. Until we have made operational these notions of the optimum incurring and repaying of international debt, it will be difficult to assess the performance of the private capital markets in dealing with the problem. Until these questions are worked out, it seems preferable for intergovernmental supplementary lending arrangements to err on the generous side. The role of the war debts in the world depression gives some notion of the global economic costs of mismanagement, while the role of war debts in the German depression should provide a flavor of the potential social and political consequences of excessively rapid real economic adjustment.

# THE FLEXIBILITY OF WAGES AND PRICES

## Inflation Theory and Policy

*By* William D. Nordhaus*

The short-run behavior of aggregate prices is one of the most significant and controversial of macroeconomic variables. In a much-criticized section of the *General Theory*, Keynes suggested that a capitalist economy in recession would show little price response to changes in aggregate demand. At the other extreme, some monetarists have argued that changes in aggregate demand lead rapidly to price changes. Phillips and his followers lie somewhere between these views, holding that some but not all of an increase in nominal demand ends up in higher prices. Where does this debate stand today? My reading is that, while far from the truth 40 years ago, the truth is coming closer to Keynes every year. The current state of the *U.S.* economy, and particularly the refusal of wage inflation to moderate significantly in recession, demonstrates how much inflation has been insulated from the pressure of demand.

Much theoretical writing has been devoted to explaining the inflationary process. A useful way to model a Western economy resembles a theory of a "dual-economy." One set of markets, call them *auction markets*, are the typical competitive supply and demand model, exhibiting flexible prices. Sometimes these are internationally traded goods or "exposed" goods as in the Scandinavian inflation model, but this is not essential. Agricultural markets are rather clearly in this mold as are many security and raw commodity markets. At the other pole lie the *administered markets*. These tend to be markets where either buyers or sellers have significant market power, and one significant use of that power has been to restrain price movements. The administered markets contain much of the manufacturing, utility, and government sectors, and the labor markets are progressively becoming administered.

How do different market structures transmit inflation? R. G. Lipsey's original interpretation of the Phillips curve was as a Walrasian adjustment mechanism in a competitive market—wages respond to the disequilibrium in the labor market. The irony is that this seems a better description of price dynamics in an auction market, like the wheat market, than in an administered market like the labor market. A more important criticism is that the Phillips-Lipsey model has not incorporated inflation in a satisfactory manner; the customary modification of the Phillips curve for inflation is to add an *equilibrium* inflationary correction to the disequilibrium response, but this is an inelegant appendage under the Lipsey interpretation. Why should the response to a given state of disequilibrium vary at different inflation rates? This is similar to arguing that the laws of motion of a pendulum differ depending on its average velocity through space. It seems to me that the original Phillips-Lipsey curve is a reasonable way of representing disequilibrium price dynamics only in auction markets. In auction markets inflationary expectations get built into equilibrium auction-good prices through the effect on supply and demand,

* Dept. of Economics, Yale University.

and built into the time-sequence of auction-good prices through interest rates, but inflationary expectations do not get built into the disequilibrium adjustment mechanism for auction goods.

For administered markets we have a great deal of empirical evidence in hand. One area where the evidence is good concerns price behavior in the administered industrial sector. Wynne Godley, Kenneth Coutts, and I have examined price behavior in all *U.K.* manufacturing as well as in seven disaggregated industries over the postwar period. The main hypothesis we examined is whether price behavior can be adequately described as "normal pricing"—that is, are prices set on the basis of costs normalized for the cycle. We found, rather surprisingly, that prices were almost completely determined by normal costs. More precisely, once costs have been converted to normal average costs (costs are read off a cost schedule at the normal value of output) there was at most a whisper of an effect of the *actual* level of demand and costs on actual prices. Just to give an empirical flavor to these observations, we concluded that the effect of the average *U.K.* business cycle on manufacturing prices (for a given structure of costs) would be less than .1% in absolute value. It is unclear whether these limited empirical results should be extrapolated to other administered markets. Otto Eckstein and David Wyss, as well as Robert J. Gordon, have found very similar price behavior for the manufacturing and nonfood sectors of the United States, although the United States appears to be more sensitive to demand than the United Kingdom.

Do firms price labor in the same way that they price goods? Up until the time of Keynes, wages over the business cycle were assumed to be determined by the classical relation between the marginal value product and the wage. Even though Keynes later admitted that cyclical movements in the real wage were anticlassical, he did not purge this last vestige of classical thought from the *General Theory;* indeed this anachronism is still embalmed in most modern macroeconomic textbooks. A close examination of the cyclical movement in wages reveals that they parallel prices in the trend toward normal pricing. In the modern terminology, Phillips curves are apparently becoming much flatter in the short run, and it is clear that in deep recessions the marginal impact of unemployment on wage inflation is very small. At the same time, recent econometric work seems to indicate that wages are more responsive to inflation than in earlier years.

An approach to understanding these phenomena might run as follows: administered wages generally follow the principles of normal pricing, so that cyclical fluctuations are largely removed. In bargaining situations, employers are prepared to settle for a wage which will keep their profit margins constant. Obviously, this acts as a much more powerful restraint on wages when an entire industry is not at the bargaining table. Employees keep their eyes both on competing groups and on their expected growth of real income. If wages are in line with historical wage differentials, and real incomes grow as expected, after the usual rites of brinkmanship and bravado, the settlement can be reached. (In pure administered wage situations, without bargaining, the determinants would be much the same except that competing wages might be more important.)

This simplified model can be solved for the resulting wage settlement as a function of the lagged wage settlements (as well as entropic elements such as the deviation of wage differentials from a normal pattern), the rate of change of value-added prices, and the rate of change in

exogenous prices (that is, the price of imports to the administered sector). Such a bargaining model would predict that an increase in value-added price inflation would eventually be completely passed through into higher wage inflation. If the increased value-added inflation were completely balanced, then such a solution leaves all real variables (except the real return on noninterest bearing money) unchanged. But an attempt to pass through exogenous price inflation, such as that on oil and food, would be vigorously resisted. George L. Perry's evidence suggests that value-added prices do get fully passed through into wages while only 15 percent of exogenous price inflation (such as that due to food and fuel prices) is passed through.

The administered wage model points to lagged wages and value-added prices as the major fuel for continued wage inflation, while the administered price model points to normal average costs as the main determinant of prices. To complete the system, we need to consider the prices in the auction system. The principles behind auction price determinations are basically the classical principles of competitive supply and demand. The clearest examples are those for raw commodities such as those on international exchanges (copper, zinc, raw fibers, foodstuffs). Recent work in this area points to the importance of international supply and demand factors—demand appears to be at least ten, perhaps fifty times more important in explaining short-run price movements in these auction markets than it is in the administered markets. Auction prices provide much of the noisiness in price behavior over the last four years with unexplained elements — dramatic shifts in patterns of hedge buying or even speculation—playing a large role in the 1972–75 inflation.

Putting all these elements together we get three basic elements in the aggregate price equation. First, we see demand entering weakly through three routes: through the markup in the administered market price equation, through unemployment in the wage equation, and from auction markets. The nature of the nonlinearities means that all three responses are likely to be very weak when the overall level of utilization is low, but demand inflation is significant in periods of very high utilization. The second major effect is the momentum effect of the current and past rates of wage and value-added inflation in the administered price sector. Recent work indicates that, random shocks aside, the momentum effects dominate price and wage movements over the short run; this implies both that using demand to affect inflation has a small initial effect and that its effects are spread over long periods. A recent study of Michael L. Wachter, for example, indicates that the mean lag of demand on prices operating through the labor market is in the order of three to five years. The third major determinant is exogenous movements in auction prices. These may be due to a world boom largely outside the control of the national authorities, to variations in weather or ocean currents, to stock building in nonmarket economies, or to speculative fever.

The model of inflation outlined here is hardly new; Gardner Ackley outlined a similar model 15 years ago. Such a model explains relatively well price movements over the last four years (as is shown by Gordon or by John B. Shoven and myself). Why then have economic forecasts been so far off the mark in recent years? There are basically four reasons for the very poor price forecasts. First, the auction sector has been largely ignored, for differential movements in auction goods prices were insignificant until recently. Second, there appear to have been struc-

tural shifts—Phillips curves estimated today look remarkably different from those estimated five years ago. Third, even with today's models, inflation cannot be predicted with great precision. The exogenous element in auction prices was very large over the last four years, and one recent attempt to predict commodity prices shows errors of up to 50 percent in 1973 and 1974. Finally, the normal timing of the price-wage system has been wrenched by a successive tightening and loosening of price and wage controls. Gordon estimates that the loosening of controls in 1974 led to an inflation rate three percent above what would otherwise have occurred.

While the exact details of price behavior may differ from that painted here, the broad vision seems to accord with much recent empirical work. Unfortunately, it is quite difficult to describe a theory which can underpin the reality—leading to the description of administered pricing as a phenomenon in search of a theory. After three decades, the major intellectual problem continues to be the fact that so little of the response to demand shifts comes through prices and wages. Although some economists continue to argue that price movements are underestimated because of errors of measurement (through price shading in recession and vice versa), there is little evidence that the errors are substantial. Indeed the aggregate transactions price index of George J. Stigler and James K. Kindahl shows *less* cyclical sensitivity than the corresponding official price index. Moreover, the pervasiveness of price stickiness across time and space casts doubt on behavioral explanations that rely on the weakness of forces of selection to drive out nonoptimizing behavior without showing why such nonoptimizing behavior arises in so many situations in the first place.

Early explanations of normal pricing (such as Paul M. Sweezy or the more elegant Edmund Phelps-Sidney Winter model) relied on the structure of competitive forces. In Phelps-Winter, for example, firms respond with quantity because customers will emigrate to other firms if they do not. This class of explanation fails to explain why prices respond so much more completely and rapidly to cyclical changes in costs than to cyclical changes in demand. A more recent rationale elaborated by Arthur M. Okun embeds wage and price behavior in a search theoretic framework. Price stability on the part of a firm is seen as part of a more general trend toward standardization of product; when I drive into a MacDonalds, I know not only the menu, quality, and speed of service, but also the prices. The "normal" pricing rule can be interpreted as a form of implicit long-term contract, where the consumer is willing to pay slightly more, averaged over the cycle, for such a standardized product because of the reduction in necessary search costs. Why is normal pricing the standardized product that so many firms sell? A normal pricing rule is an equitable-sounding and easily-understood rule. Such a rule might well be the solution to an optimization problem in which the variance of prices is minimized subject to an average profits constraint. The new view of sticky prices as a service is a revolutionary reversal of populist views of Gardiner C. Means and his followers. Unfortunately, the new view has not been tested (nor are refutable hypotheses clearly stated).

The "dual economy" view of inflation has important implications for macroeconomic theory and policy. Thus the empirical description of the inflation transmission process casts doubt on extreme neoclassical and monetarist views of inflation. One such contention is that inflation is everywhere a monetary phenomenon. Presumably, in plain English, this means

that the money supply exerts a direct influence on the evolution of prices. In analyzing this question, one must be extremely careful to separate out the reduced-form effect from the structural effect. In the model outlined in the present paper, and neglecting second-order effects, money affects the price level only through its effect on aggregate demand. Thus in the normal price model, money does not enter directly into the structural equation but enters only through factor prices, while in the auction market money enters only by affecting excess demand. On the other hand, in a reduced form equation, all modern models except the extreme Keynesian depression models have money as an important determinant of aggregate demand, which influences demand curves in the auction markets and unemployment in the Phillips curve. Thus money should be insignificant in a properly specified structural price equation, while in a reduced form price equation money should be a powerful variable.

The distinction between reduced form and structural equations helps explain much recent econometric work. Reduced form price equations with money have performed surprisingly well. Yet in structural equations, such as in the Nordhaus-Godley, there is no perceptible influence of reversible cyclical factors in price determination, and there is no room for money to exert a direct role in price determination in the administered price sector. Gordon has found that a bivariate reduced form price equation with money is highly significant with a very long lag (much longer than is consistent with popular monetarist writings), but that adding money to the structural equation shows very little independent influence of money outside of the influence on the structural variables.

A recent strand of neoclassical thinking, associated with the writings of Robert Lucas, Thomas J. Sargent, and Neil Wallace, questions the efficacy of short-run macroeconomic policy because of the view of price dynamics. If expectations are rational and if prices and wages are flexible, policy is unable to "fool" people into allowing a deviation of unemployment or demand from the equilibrium rates even in the short run. These theories allow only a "learning lag" between policy and complete price flexibility. (The learning lag represents the necessary time for producers or consumers to learn that the real variables such as unemployment or output have changed.) In light of the fact that many policies are announced in advance—so that learning lags must be brief—and that the lags of output and prices in response to policy are so long, it is hard to understand how these theories could be seen as more than intellectual exercises. Put differently, is it plausible that the fall in capacity utilization over the current recession is either a fall in the "natural rate of utilization" or that it is unknown to the producers who actually leave the capacity idle?

A third major implication of the dual economy model is for international trade. Some recent work has stressed the importance of internationally determined prices in the transmission of inflation over national borders. According to the "law of one price," with fixed exchange rates, the world price level determines the price level in individual countries directly. This mechanism explains the puzzle as to the simultaneous outbreak of inflation in so many countries. Some monetarists have even pushed this to an extreme by arguing that price flexibility in individual countries will be a perfect substitute for floating exchange rates.

In terms of the dual economy model, these views seem to argue that the auction market is the dominant sector, and further that prices in the auction market are

set in world markets. Although no careful study of this proposition has been made, it seems clear that the significance of the "law of one price" is overemphasized. Clearly labor does not have one price, and outside of true auction markets there appears to be sufficient product differentiation and transport and tariff costs so that world prices exert but little influence on domestic prices in the short run. Godley, Coutts, and I have examined seven manufacturing industries in the United Kingdom to see whether the world price exerted any effect as against the normal pricing rule. We found on average the effect of the price of competing imports in the same industry was about one-tenth as strong as would be predicted by the law of one price. Similarly, Gordon found no room for competing export or import prices in his study of aggregate price behavior in the United States. I believe that these results cast suspicion on recent theories which rely on rapid adjustment of prices to either world prices or to demand.

The dilemma for policymakers in choosing between inflation and output is not only cruel but becoming crueler. Indeed, the inability to control both inflation and unemployment within acceptable institutions is the major flaw of Western economies today. The aversion to inflation has led Western countries into a recession of huge proportions, yet the inflation has hardly abated. Given the long lags in response, one must marvel at the steely nerves of politicians who sacrifice 10 percent of output today for a point or two in the inflation rate spread over the next five years. But in the long haul, there is a tension in a system where democratically-elected policymakers must choose between today's output and employment and next year's inflation. This kind of choice has led to political business cycles in the past and is embodied in the 1975 *U.S.* energy bill where politicians would only bite the bullet of price decontrol as long as the price explosion would come after November 1976.

Given the increasingly cruel dilemma and the unsatisfactory response of policymakers to the dilemma, it is tempting to search for new ways of controlling inflation. The ideal form of inflation control is one which reduces all nominally denominated values by the same percent without changing any relative prices or real magnitudes. The Achilles heel of past price-wage policies (except the relatively innocuous guideposts) has been that they confused inflation policy with income redistribution. By concentrating on individual sectors, they have attempted to change relative prices and aroused the fierce resentment of individual groups. Whether the 1962 steel controversy was justified or not, it can hardly be seen as distributionally neutral. The current contractionary route to deflation has even graver distributional consequences.

There is probably no such ideal anti-inflation policy, but economists have shown little inventiveness in designing durable antidotes to inflation other than recessions. One serious suggestion is an inflation tax which would penalize firms or workers to the extent that they deviated from a national norm. Such a mechanism would allow the decentralized decisions of which economists (and some politicians) are so fond. It would allow adjustment of relative prices, and when inflation rate is at the norm it would leave relative prices unchanged. Most important, such a measure would directly affect the bad—inflation—rather than first pummeling goods like output and employment.

# Some Problems in Wage Stabilization

By MICHAEL L. WACHTER*

A major issue in inflation theory and in the related empirical policy questions is the ability of the government to control wage inflation, in the short or the long run, through fiscal and monetary policy. This paper discusses issues involved in this question and offers some tentative conclusions. The problem areas can be divided into three headings: the difficulties in measuring a noninflationary rate of unemployment, the sources of wage rigidity in the labor market, and the impact of cost-push wage inflation.

## I. The Noninflationary Unemployment Rate

Attempts to stabilize the wage inflation rate depend on the ability to measure either the tradeoff between wages and unemployment, i.e., the Phillips relationship or to identify the unemployment rate (or rate band) beyond which the inflation rate tends to accelerate. Until recently most of the empirical literature focused almost exclusively on estimating the short-run Phillips curve.This functional relationship, however, proved itself an unstable tool for stabilization policy. The most widely accepted explanation for this instability is found in the various accelerationist models of inflation. In these models attention in stabilization policy is focused on the noninflationary rate of unemployment.

The natural rate or full employment rate, as originally discussed by Milton Friedman, referred to the long-run, "market clearing" unemployment rate in the labor market. At this rate, the actual rate of wage inflation is equal to the expected rate of wage inflation, and the economy has a tendency to gravitate towards this level of unemployment if left to its own devices. In addition, the natural rate might be expected to change only slowly over time. If the natural rate is known, stabilization policy can be utilized to speed a recovery when shocks in the private sector force a divergence from full employment.

An important development since the appearance of the natural rate literature is the notion that the unemployment rate at which inflation appears to be unchanging is not or at least has not been stable over the recent past and is not known with any degree of confidence. Specifically, the noninflationary unemployment rate ($U_{NI}$) is a short-run phenomenon incorporating an array of labor market forces. Some of the arguments of the $U_{NI}$ function, such as secular changes in relative wages or prices, are compatible with the natural rate concept of Friedman; other factors, are essentially nonmarket clearing, disequilibrium events. Incorporating this broad array of factors, it is not surprising that $U_{NI}$ is very difficult to calculate. In this section I discuss a few of the potentially large number of arguments of $U_{NI}$.

The evidence suggests that several independent and related forces have been operating since the early 1960's to push up $U_{NI}$ quite sharply. Of particular importance is the demographic shift towards younger workers and female workers. (See George Perry.) This is a reflection of the baby boom generation reaching the labor

* Associate professor of economics, University of Pennsylvania. This research was supported by the National Science Foundation and the National Institute for Health. The author wishes to thank Arthur Okun, Thomas Sargent, and Susan Wachter for many helpful comments.

market as well as the increase in the labor force participation rate of women and especially of younger women. Young workers, who naturally have relatively little specific human capital, have historically always had the lowest wages and the highest turnover rates. This shift in the labor force essentially suggests measuring workers in efficiency units.

A second force which may have been increasing $U_{NI}$ over the recent past is the rise, abstracting from cyclical swings, in the unemployment rates for the young and most female demographic groups. The basis for this position is that different skill groups are imperfect substitutes for each other so that a significant relative abundance of one group should cause an alteration in relative wage differentials. But if wage differentials among demographic groups are not sufficiently flexible, markets will not clear and unemployment rates will be forced to change as well (or in place of relative wages). In fact, the adjustment in a world where the skill requirements for capital equipment, etc., are largely fixed is likely to make it difficult for relative wages to clear the market. Further weakening the tendency of relative wages to change has been government policy measures. Dramatic changes in minimum wage coverage have prevented demand adjustments favoring lower skilled workers, and changes in unemployment compensation and welfare have steadily increased the *relative* reservation price of labor thereby lowering the cost of being unemployed. Studies of these policy measures, on balance, strongly support a displacement effect that would serve to increase $U_{NI}$.[1]

My attempts to date at creating a $U_{NI}$ series have concentrated on the demographic shifts described above and, to a lesser extent, the changing cost of unemployment. The estimated $U_{NI}$ series is shown in Figure 1. Given data and methodological problems, the series is a crude measure and, in particular, smoothes out much of the short-run variation in $U_{NI}$. The current estimated level of $U_{NI}$ in Figure 1 is 5.5 percent. Reasonably small changes in the method of constructing $U_{NI}$, however, can alter the estimated current level to as high as 6 percent.[2]

The above discussion, however, does not exhaust the potential variables which may be arguments of the $U_{NI}$ function. For example, the secular (as distinct from cyclical) variation in the wage structure may be positively related to $U_{NI}$. If all firms were to pay identical wages (adjusting for worker quality), workers would accept the first job opening and would have little incentive to wait for a "better" job. If, for a variety of institutional reasons, some firms pay higher wages than others, workers may prefer to be "frictionally" unemployed on the queue of the high wage sector. (This unemployed group includes those on lay-off from the high wage sector.) The postwar widening of the wage structure makes it increasingly profitable for some workers to remain unemployed on the queue of the high wage sector rather than accepting employment in the low wage sector, thus increasing $U_{NI}$. (See, for example, Robert Hall and Stephen Ross and Michael Wachter.)[3]

[1] See, for example, Martin Feldstein, Finis Welch, and Jacob Mincer. The impact of changes in the relative population of young workers on their income is discussed in Richard A. Easterlin.

[2] The method for calculating $U_{NI}$ is described in detail in M. Wachter (1975). If one accepts the notion that $U_{NI}$ is stable only in the short run, then its current high level need not persist. The demographic trends are favorable in that the percentage of teenagers in the population has already peaked and the average worker will be moving into an age category which typically has less mobility and long job attachments. My projections using fixed participation rates suggest a decline in $U_{NI}$ over the next ten years—slightly greater than its rise over the preceding ten years—to approximately 4.5 percent.

[3] A very different approach to the issue of identifying $U_{NI}$, which also stresses its unstable character,

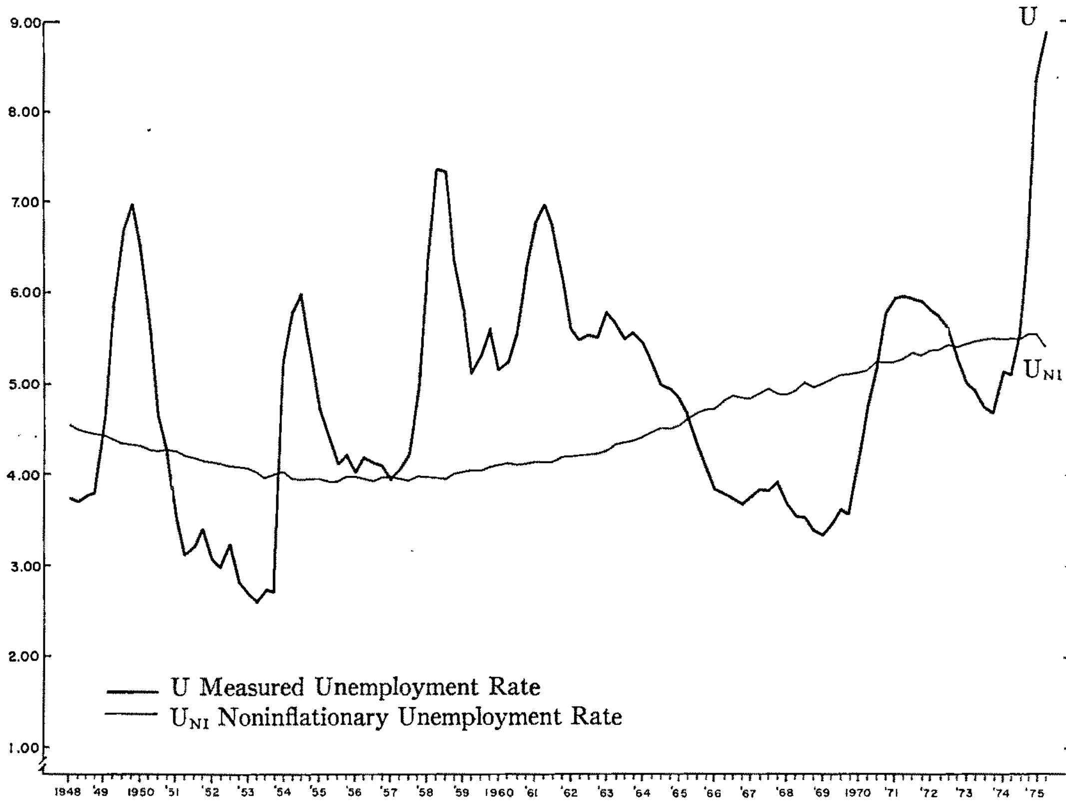



FIGURE 1

In summary, then, if the natural or the noninflationary unemployment rate were constant over time, there would be few problems in estimating the rate on the basis of time series wage data. However, it appears not only that $U_{NI}$ is a variable, but also that it is difficult to predict. The uncertainty over the level of $U_{NI}$ represents a serious problem facing stabilization policy.[4] In fact, one can argue that the major cause of the current inflation was a miscalculation of $U_{NI}$ in 1967–69 and again in 1972–73. Specifically, during 1967–69 the trade-off view was almost uniformly accepted, with the result that an unemployment rate well below 4 percent was adopted as at least a short-term target. The cost was presumed to be a somewhat higher but stable rate of inflation. In 1972–73, on the basis of government policy and pronouncements, it seems likely that the $U_{NI}$ rate was assumed to be approximately 4 percent rather than the

appears in some of the recent rational expectations models. In the Thomas Sargent and Neil Wallace model, quarterly movements in $U_{NI}$ may be calculated from an autoregression of $U_t$ and past values of $U_{t-i}$. That is, $U_{NI}$ is only separated from measured unemployment by an error term. This defines $U_{NI}$ in a purely statistical manner and does not attempt to attribute particular economic causation to the variation in $U_{NI}$. In this "superneoclassical" framework systematic monetary and fiscal policy are powerless to alter the unemployment rate. The reason is that the short-run Phillips curve (as well as the long-run curve) is vertical. Hence, only "surprises" in policy can alter the measured unemployment rate.

[4] This brief summary has omitted the possibility that even in the long run the Phillips curve is not vertical, hence $U_{NI}$ is not a single number at any point in time. Nonvertical long-run trade-offs appear in James Tobin, where stochastic disturbance leads to constant market disequilibrium, and Stephen A. Ross and Wachter, where institutional contracts are not perfectly indexed on inflation.

5 to 5.5 percent figure indicated in Figure 1.

## II. Wage Rigidity

A second difficulty facing stabilization policy is the inflexible response of at least some wages in the economy to unemployment. Inflexibility arises from a number of factors. The most widely accepted is based on the behavior of unions where collective bargaining introduces obvious lags in the wage response pattern of affected industries. Alternatively, firms often desire to pay a wage premium even in the absence of a union. The premium gives a preferred position in the labor market, assuring the firm of a queue of available workers, and enables them to avoid altering wages immediately as labor market conditions change.

Recent models have also attributed wage rigidity to the desired ongoing relationship between firms and their work forces.[5] For example, this may arise when the job content of the firm is idiosyncratic as a result of job specific training. As a consequence, most jobs in these high wage firms do not have a direct demand and supply component external to the firm. Rather, the jobs are part of the internal labor market of the firm where they are connected through a series of promotion ladders which have their own rules to provide enforcement and information on how the "market" works. Wages on each job are set as part of the internal wage structure of the firm. External market wages do not set any individual firm wages; rather the external market impinges on the firm by influencing the availability of new workers and the quit rate of old workers. Promotions and specific training, rewarded by the internal wage structure, mean that the opportunity wage of workers is below their current wage and this discourages mobility. Firms are reluctant to try to capture the discrepancy, lest it encourage costly mobility of trained workers and because of the need to encourage workers to utilize and learn the specific (and general) knowledge needed for different positions along the promotion ladder. The ongoing relationship results in firms paying a wage premium above the opportunity wage of their specifically trained workers. This premium gives firms the ability to ignore short-run market forces.

In all of the above models, wage premiums and the resulting wage rigidity arise because of the existence of institutions based on long-run commitments which are not responsive to short-run economic stimuli. As distinct from neoclassical search models, the rigidity is not based on the problems that rational workers and firms have in trying to guess the present and near term nominal and real wage levels. Rather, it is due to the fact that firms and workers cannot have all of the necessary information on the future states of the world, including fluctuations in aggregate demand policy, when they establish their contractual and institutional arrangements. Indeed, the very purpose of these institutions is to allow firms and workers to cooperate without having to sign contracts which would detail their responses to all of the potential future contingencies that might arise. These models may well be rational, in the current use of that term, but only over the long horizon during which full recontracting occurs.

The institutional nature of these models has led some to suggest that wages are not just inflexible, but rather are not determined by demand factors. For example, unions are influenced by the settlements won by other unions as part of a cost-push

[5] There is a growing literature in this area. See the summary statement by Robert J. Gordon. The discussion in the text follows Okun and Oliver Williamson, Wachter, and Jeffrey Harris.

process. This seems to be supported, in part, by recent empirical results on wage-price equations. The equations appear to show that wages depend largely on past prices and/or wages and only slightly on unemployment. Prices, in turn, depend upon cost variables. Since the cost variables are essentially prices (with wages being the most significant single price) the system reduces to a highly autoregressive model with unemployment or demand seemingly having a minor role to play.[6]

My argument is that this view of the inflation process is based on a very narrow interpretation of the data and is not in keeping with the underlying causes of wage rigidity discussed above. Even in the traditional structural wage equation where wage inflation is a function of unemployment and lagged price (or wage) inflation, the autoregressive price term should be viewed as a distributed lag generator of past conditions in the labor market. That is, lagged prices reflect expectational effects, or more generally, inertia effects in the labor market. For example, in the unionized sectors and in the internal labor markets where wages are set at discrete intervals, the economy-wide price or wage history over the fixed period of individual firms provides those firms with information on labor market conditions between their own wage changes. As a result, the feedback response through the lagged price (or wage) variable is central to the question of the magnitude and speed with which wage inflation responds to demand variables. Recessions alter inflation not only by moving the economy along a given Phillips curve, but also by causing that Phillips curve to shift downward over time. Indeed, it is difficult to separate, on both a theoretical and an empirical level, movements along and shifts of the Phillips relationship. Where the direct unemployment effect ends and the induced shift of the curve begins is largely a definitional question.

Supporting evidence for this interpretation can be found by replacing the distributed lag on prices with a distributed lag on money supply changes. The two equations have similar overall statistical fits. However, whereas lagged prices can be interpreted in a purely autoregressive framework, the money supply term more clearly reflects aggregate demand forces. Hence, one can estimate a quasi-reduced form Phillips curve which contains the discrepancy between measured unemployment and $U_{NI}$ and a distributed lag on money growth. The results indicate a short-run downward sloping Phillips relationship and a long-run wage inflation which is homogeneous of degree one in money growth. (More specifically, the long-run coefficient on the money supply is below unity but not significantly below unity. The same is true for the long-run coefficient on the price term.)

The conceptual problems in the long lags are reasonably simple, but they do create difficult policy situations. In particular, the political goal of significantly less inflation cannot be achieved quickly, and, indeed, the wage inflation rate may act in a perverse manner in the short run, depending upon the structure of the lags and the recent heritage of price inflation. That is, even though wage inflation is ultimately a function of aggregate demand,

[6] Those who espouse the nearly horizontal Phillips curve or the exogenous inflation rate also tend to be those who believe that the full employment, unemployment rate is considerably below the rate indicated in Figure 1. If one views $U_{NI}$ as being low, perhaps 4 percent, then the inflation since 1970 must largely be due to exogenous elements. As a consequence of long lags, some of today's inflation could certainly be due to the period preceding 1970 when unemployment rates were below 4 percent. But with unemployment above 4.5 percent since 1970, it is difficult to explain the current high rate of inflation without believing that $U_{NI}$ is greater than 4.5 percent *or* that inflation is an autoregressive process independent of excess demand.

much of the short-run variation in inflation is due to the past history of excess demand and not to the current unemployment rate. This creates a political dilemma as was seen in 1971 and may well appear again in 1976. In the midst of a period of high or rising unemployment, the slow downward movement of the wage inflation is very unwelcome. It does, however, correctly signal the nature of the problem. The wage adjustment is slow because demand factors take a long time to work completely through the complex institutional arrangements in the labor and goods markets. It is, however, lagged demand that is locked into the heritage or inertia effects, and these forces cannot be repealed by simply relabeling them as predetermind lagged price effects.

### III. Cost-push Inflation

The third element in the current debate over wage stabilization policy concerns the role of cost-push wage inflation. This is closely related to the previous issue of the exogeneity of the rate of wage inflation. In a sense, cost-push inflation may be viewed as one economic explanation for the autoregressive view. Unfortunately, cost-push inflation has often been defined in the past as simply the wage increases which occur when the economy is in a recession. That is, "costs" rather than "demand" are pushing up wages. If, however, the supposed cost-push phase is simply the lagged demand-pull forces working in the high wage or slow response sector, this definition is not useful in that it is devoid of policy implications.

As I have suggested elsewhere (1974), an alternative definition of cost-push can be based on changes in the relative wage structure. Specifically, the offending cost-push sectors would be those that have been secularly increasing their wage differentials, after correcting for the predictable cyclical fluctuations in the wage structure. In this definition, cost-push is not simply a lagged manifestation of demand-pull inflation, but is rather a nondemand or "exogenous" element. The union-oligopoly sector in manufacturing, which is most often labeled as cost-push, exhibits a clear cyclical pattern in its wage differential over the competitive sector but little secular rise. There are several sectors of the economy, however, which have for some time exhibited rising secular wage differentials. The two long-standing candidates are the various sectors of the government, federal, state, and local, and the construction industry. A more recent entry is the transportation and public utilities sector.[7]

The presence of cost-push inflation, defined by an increasing wage premium in certain sectors, is to reduce the attractiveness of aggregate demand policies. First, as mentioned above, cost-push may increase $U_{NI}$. Secondly, it will have a direct wage effect which may appear as an increase in the constant term or the coefficient of the autoregressive term of the short-run curve. The magnitude of the cost-push problem, however, is easily overstated. Although an important proportion of the economy is involved in the process, I argue that only their wage premium increases should be included.[8]

This sector approach to the cost-push issue also suggests an additional policy instrument for wage stabilization which can take some of the social cost out of the "recession cure." The cost-push sectors in the *U.S.* economy have been troublesome

[7] See Robert E. Hall for a more detailed discussion of the wage problem in what he calls the "nonentrepreneurial sector."

[8] To measure the full impact of cost-push pressures necessitates some method for including spillover effects. A more important issue, however, is that ongoing cost-push inflation must essentially be ratified by the monetary authorities and hence is indistinguishable from traditional demand inflation. Due to space limitations other types of exogenous shocks to the wage process, besides cost-push factors, are not discussed.

for some time. The reason is that they have highly inelastic labor demand functions, in part due to "favorable" government legislation which affects these industries directly or indirectly. They have thus been able to afford an increasing wage premium with little effect on employment. Although there has recently been some erosion of government employment protection in the state and local government sectors and in construction, the longer term trends are still unclear. What is needed in the cost-push labor markets are sector-specific, structural programs which will augment competitive demand forces.

## IV. Conclusion

To summarize, I have discussed the impact of three potential problems that confront policy makers in attempting to control wage inflation. First, the difficulty in estimating the noninflationary unemployment rate, is a very real problem. In fact, I would argue that incorrect $U_{NI}$ targets over the past decade explain an important part of the current inflation problem. Secondly, wage inflexibility is essentially a political rather than an economic problem. Wage inflation does respond to aggregate demand but with reasonably long lags due to institutional arrangements in product and labor markets. Thirdly, cost-push inflation is generally ill-defined so that it is difficult to classify. I argue, however, that it is best defined and dealt with on a sector-specific basis. As such, it is not a problem for aggregate demand management.

## REFERENCES

**R. D. Easterlin,** *Labor Force and Long Swings in Economic Growth: The American Experience,* New York 1968.

**M. S. Feldstein,** *Lowering the Permanent Rate of Unemployment,* A Study Prepared for the Use of the Joint Economic Committee, 93 Cong. 1 Sess. 1973.

**M. Friedman,** "The Role of Monetary Policy," *Amer. Econ. Rev.,* Mar. 1968, *58,* 1–17.

**R. J. Gordon,** "Recent Developments in the Theory of Inflation and Unemployment," paper presented at the Conference on Inflation Theory and Anti-Inflation Policy, Sweden, Aug. 1975.

**R. E. Hall,** "The Rigidity of Wages and the Persistence of Unemployment," *Brookings Papers,* 1975, *3,* 301–35.

**J. Mincer,** "Unemployment Effects of Minimum Wages," forthcoming, *J. Polit. Econ.,* 1976, *84.*

**A. M. Okun,** "Inflation: Its Mechanics and Welfare Costs," *Brookings Papers,* 1975, *3,* 351–90.

**G. L. Perry,** "Changing Labor Markets and Inflation," *Brookings Papers,* 1970, *3,* 411–41.

**S. A. Ross and M. L. Wachter,** "Wage Determination, Inflation and the Industrial Structure," *Amer. Econ. Rev.,* Sept. 1973, *63,* 675–92.

**T. J. Sargent and N. Wallace,** "Rational Expectations, the Optimal Monetary Instrument, and the Optimal Money Supply Rule," *J. Polit. Econ.,* Apr. 1975, *83,* 241–54.

**J. Tobin,** "Inflation and Unemployment," *Amer. Econ. Rev.,* Mar. 1972, *62,* 1–18.

**M. L. Wachter,** "Phase II, Cost-Push Inflation, and Relative Wages," *Amer. Econ. Rev.,* June 1974, *64,* 482–91.

———, "The Changing Cyclical Responsiveness of Wage Inflation over the Postwar Period," University of Pennsylvania mimeo, 1975.

**F. Welch,** "Minimum Wage Legislation in the United States," *Econ. Inquiry,* Sept. 1974, *12,* 285–318.

**O. E. Williamson, M. L. Wachter, and J. E. Harris,** "Understanding the Employment Relation: The Analysis of Idiosyncratic Exchange," *Bell J. Econ.,* Spring 1975, *6,* 250–78.

# Inflationary Tales Told by Static Models: The Case of Price Setters

*By* GEORGE A. AKERLOF*

Macroeconomic analysis has rediscovered the disequilibrial method of the Swedish economists. According to this analysis *ex ante* decisions are made in period t to be carried out in the next period t+1. In most models using the disequilibrial method, *ex ante* decisions have only concerned quantities. Yet, as Swedish economists were aware in the 1930's (see E. Lindahl) and as is implicit in all econometric models with wage and price equations, in a world of price-setters these decisions at time t will concern prices as well as quantities at time t+1.

The disequilibrial method is particularly easy to apply to prices. Since the *ex ante* decisions of different economic agents may be mutually incompatible (for example, *ex ante* savings may not equal *ex ante* investment), disequilibrial models must also specify the mechanism whereby the potentially conflicting *ex ante* decisions are translated into *ex post* reality. Typically, no single mechanism is obviously superior to all others. For example, the usual Keynesian model, in which unintended inventory accumulation is the difference between *ex ante* savings and investment, is only one of several equally plausible models. In contrast, there is a single most natural relation between *ex ante* and *ex post* prices: that the *ex post* nominal price equal the *ex ante* nominal price.

The equivalence of *ex ante* and *ex post* nominal prices does not mean that all price setters are satisfied *ex post*. On the contrary, given a rational world, economic agents make plans regarding relative prices; since the agent controls his own price, but not the price of other agents, he may still be disappointed *ex post* with relative prices, despite equality of *ex ante* and *ex post* nominal prices. This disappointment, by our theory, is the spur to continued inflation, given constant demand and constant expectations regarding the rate of inflation.

## I. Ex Post Inflation and Ex Ante Price Behavior

Let $\psi$ be a price index. $\psi$ is by definition a function, homogeneous of degree one, $\psi_t=\psi(p_{1,t}, \cdots, p_{n,t})$ where $p_{i,t}$ is the price of good $i$ at time t. Let there be an expected rate of inflation, denoted $\gamma^e$, which is uniform for all economic agents. Suppose that at time t the *ex ante* real price of good $i$ is $(p_i/\psi)_t^D$. The equality of *ex post* and *ex ante* nominal prices yields:

$$(1) \qquad p_{i,t+1}/\psi_{t+1}^e = (p_i/\psi)_t^D$$

where superscript $e$ denotes "expected"; and then

$$(2) \qquad \psi_{t+1}/\psi_t = \psi(p_{1,t+1}, \ldots, p_{n,t+1})/\psi_t$$

$$(3) \qquad = \psi((p_1/\psi)_t^D, \ldots, (p_n/\psi)_t^D) \cdot (1+\gamma^e).$$

In simple Keynesian economics the fundamental equation relates *ex post* income to *ex ante* investment and savings decisions. Equation (3) shows how *ex post* inflation is related to *ex ante* pricing decisions.

The theory of inflation is complete with determination of the *ex ante* decisions

* Associate Professor, University of California, Berkeley, and Visiting Research Associate, Federal Reserve Bank of New York. The author would like to thank Karl Shell, David Cass, Bent Hansen, and especially Thomas Rothenberg for valuable comments.

$(p_i/\psi)^D_t$. Each decision is pictured as the outcome of a bargain made by a small group of collusive agents. Tracing through the variables exogenous to this bargain reveals the dependency of the $(p_i/\psi)^D_t$.

The group of colluders, which may set the prices only of goods, or only of factors, or of a mixture of the two, collusively choose their expected prices so as to maximize a function $B$ of their individual utilities $U_i$. Taken over the range of possible functions $B$ and utility functions $U_i$, this specification embraces most standard bargaining solutions.

According to standard assumptions, $U_i$ is pictured as dependent on the quantity of good or factor $i$ sold by agent $i$, on the real price received, and on its real cost of production (if any)—with demand, in turn, dependent on real prices, on parameters describing tastes and tax structure and on the three variables summarizing government macropolicy (real balances $M/\psi$, real government expenditure $G/\psi$, and real taxes $T/\psi$), and with cost dependent on real factor prices, on output, and on parameters describing the structure of technology and taxes.

At time t real prices are chosen to maximize the expected value of $B$ at time t+1. We deal with expectations by assuming that $B$ and its direct and indirect arguments, such as utilities, demands and costs, in turn depend on the expectations of their arguments, of which the most important, in this regard, are expected real prices of goods and factors. Because bargains are made concerning real variables, it is reasonable (although not necessary for rationality) that expectations, likewise, concern real (rather than nominal) variables. Accordingly, expected real prices, by assumption, depend upon the expected real prices of the colluders, which are endogenous; and they depend also, because of adaptive expectations and real-world stickinesses, on the real price vector at time t, which is exogenous.

Denote the parameter vector describing tastes, etc., by $\alpha$, and the real price of factor $j$ at time t by $(w_j/\psi)_t$, and recounting the variables exogenous to the maximization of $B$:

$$(4)\quad (p_i/\psi)^D_t = F_i(\ldots, (p_k/\psi)_t, \ldots, (w_j/\psi)_t, \ldots; (M/\psi)^e, (G/\psi)^e, (T/\psi)^e; \alpha).$$

Since there is no formal distinction between bargains involving goods and bargains involving factors, *ex ante* real factor prices $(w_j/\psi)^D_t$ are functions of the same arguments.

Equations (3) and (4) divide the study of inflation into three parts: the first is the functions $F_i$, the second is the behavior of relative prices and demand, and the third is the expectations term $\gamma^e$. Study of the incidence of exogenous changes on relative prices can, by this theory, be fed into the functions $F_i$ and weighted by $\psi$ to examine the incidence of such shifts on $\psi_{t+1}/\psi_t$. These are our inflationary tales told by static models, of which the next section gives four simple illustrations.

## II. Illustrations

### A. *Example I*

As the simplest example, consider an economy of monopolistically competitive firms which share demand according to the demand function:

$$(5)\quad D_i = (p_i/\psi)^{-\mu} X^D(M/\psi, G/\psi, T/\psi),$$

where $X^D(\cdot, \cdot, \cdot)$ is the real demand per firm associated with the macropolicy $(M/\psi, G/\psi, T/\psi)$, and the parameter $\mu$ is greater than one. Each firm has a production function for its output: $Q = K^\beta L^{1-\beta}$, where $K$ is capital and $L$ is labor. Profit maximization and monopolistic competition, with constant real wage expectations and capital considered a fixed factor, yields:

$$(6)\quad (p_i/\psi)^D = [\{(w/\psi)\mu(\mu-1)^{-1} \cdot (1-\beta)^{-1}\}^{1-\beta}(X^D/K)^{\beta}]^{1/(\beta\mu+1-\mu)}.$$

In this economy, (6) multiplied by $(1+\gamma^e)$ gives $\psi_{t+1}/\psi_t$ by (3), since all firms are the same.

Formula (6) illustrates in particular what formulas (3) and (4), taken together, suggest more generally. Formula (4) lists the factors involved in the desired readjustment of relative prices: demand and relative prices themselves. Formula (6) suggests that policies with high incidence on real factor prices typically lead to reduced inflation—demand and $\gamma^e$ being constant.

As one application, formula (6) and others similarly derived predict how much inflation will occur if price controls are imposed and then suddenly removed. The inflation will be positively related to $(w/\psi)$, $X^D$ and $\gamma^e$, and negatively related to $K$. The old-wives' tale is commonly believed that wage and price controls inevitably lead to suppressed inflation, which, on removal of the controls, results in a fresh inflationary outbreak; formula (6) multiplied by $(1+\gamma^e)$ indicates the degree of suppressed inflation as well as its functional dependence.

### B. *Example II*

(1) *Question.* Tax collections being of such large magnitude, shifts in tax structure, if they affect costs but not demand, constitute a practical means of reducing inflation while keeping employment constant. The next three examples illustrate the type of tax shifts which reduce inflation. The first example concerns an upward shift in the income tax and a downward shift in the sales tax of equal amount, so that disposable income, and therefore, demand, is left constant.

(2) *Model.* Let the economy be composed of similar firms, with demand for each firm depending on the price to the consumer $p$, the prices charged by other firms $\psi$, real balances $M/\psi$, real government expenditures $G/\psi$, and real taxes $T/\psi$. The firm has a real cost function, dependent on the cost of hiring labor (which is dependent on the real wage denoted $\omega$), the real cost of materials obtained from other firms, equal to 1, the real price of raw materials (denoted $\rho$), and the sales tax on output at rate $\tau_s$. (Taxes enter the cost function since they are as much a cost of selling the product as the cost of any input.) Furthermore, firms have static expectations regarding real factor costs.

Entering the relevant variables into equation (4), multiplying by $(1+\gamma^e)$, and putting the product in differential form for convenience of notation, yields:

$$(7)\quad d\ ln\ \psi/dt = \delta[F(\omega, \rho; M/\psi, G/\psi, T/\psi\ \ \tau_s) - 1] + \gamma^e,$$

where adjustment coefficient $\delta$ enters because the length of period in difference equation (4) was unspecified. $\rho$ is taken to be a parameter determined exogenously in international markets. To track inflation in excess of $\gamma^e$ it remains to track the path of $\omega$.

Econometricians have typically specified two forms of wage equation. In the first type, due to Phillips, the expected change in the real wage is a concave function $f$ of the rate of unemployment $u$. The second type of wage equation, due to Edwin Kuh, says that the expected change in the real wage depends upon the gap between a target real wage $\omega^*$ and the actual real wage. $\omega^*$ may depend upon real variables—the sales tax and income tax (denoted $\tau_y$) among them. There being no conclusive evidence to favor a wage equation of either one form or the other, we propose the combination:

$$(8)\quad d(ln\ (w/\psi)^e/dt = \lambda(\omega^*(\tau_s, \tau_y) - \omega) + f(u).$$

Subtraction of (7) from (8) yields the differential equation for $\omega$:

$$(9)\quad \dot{\omega}/\omega = \lambda(\omega^* - \omega) + f - \delta(F - 1).$$

(3) *Result.* Our method consists of looking at the incidence on relative prices of some exogenous shift and using equations (3) and (4) to determine the effect on inflation. The incidence of tax shift on $\omega$ between steady states with constant $\omega$ is implicitly defined by (9); inflation is given by (7).[1] Demand, and therefore $u$, is unaffected by the shift in tax structure. Therefore we calculate the shift in inflation between steady states with constant $\omega$ and with $d\tau_s = -d\tau_y$:

$$(10)\quad \frac{d(\dot{\psi}/\psi - \gamma^e)}{d\tau_s} = \delta\lambda\left(\frac{\partial F}{\partial \tau_s} - \frac{\partial F}{\partial \omega}\frac{\partial \omega^*}{\partial \tau_y} + \frac{\partial F}{\partial \omega}\frac{\partial \omega^*}{\partial \tau_s}\right) \Big/ \left(\delta\frac{\partial F}{\partial \omega} + \lambda\right).$$

Interpret $\omega^*$ as the *ex ante* real wage. Formula (10) says that the change in inflation due to the shift in tax structure has the same sign as the increase in *ex ante* price directly due to tax changes ($\partial F/\partial\tau_s$) plus the increase in *ex ante* factor prices due to the tax changes ($-\partial\omega^*/\partial\tau_y$ and $\partial\omega^*/\partial\tau_s$) weighted by their effects on *ex ante* prices (weight$=\partial F/\partial\omega$).

Formula (10) gives a sensible result regarding the effect of the decrease in the sales tax *cum* increase in income tax. Remember that $\omega$ is the before-tax wage to the worker divided by the after-tax price. If the increase in income tax causes an equivalent increase in *ex ante* wages $\omega^*$ and if the pass-through of wage costs and sales taxes are equal, the inflationary effect is nil (assuming also $\partial\omega^*/\partial\tau_s=0$). The incidence of the tax shift on *ex ante* real factor cost determines whether the shift in tax structure is inflationary or deflationary. Thus our example illustrates the earlier assertion that shifts in structure tend to be inflationary or deflationary as their incidence on factor costs is positive or negative. (In other problems demand may shift, and it also must be considered.)

Formula (10) also shows the type of information, with this model, needed to predict changes in inflation due to tax structure. Most important is the effect of tax changes on intended bargains to know whether the shift in tax structure is inflationary or deflationary. Professional labor arbitrators, whose job is to know how different variables affect intended contracts, must, at least on a case-by-case basis, have informed opinions regarding the extent to which intended bargains concern before-tax or after-tax wages.

### C. *Example III*

To examine further the relation between inflation and tax structure, consider the impact of an equal fall in the sales tax and rise in the real cost of a raw material input used in fixed proportions. In the previous example, since the cost function is symmetric in the sales tax and the cost of such raw materials, and since it should be a matter of indifference to the wage equation whether additional payments are made to governments or to producers of raw materials, the equal and opposite changes in $\tau_s$ and in $\rho$ do not affect inflation through costs, but only through their unequal effects on demand. The raw materials producers, whether domestic or foreign, will increase their demand because $\rho$ has risen, whereas the fall in $\tau_s$ will result in no compensating reduction in government demand. (To other producers and to factors of production, the direct effects on income of the rise in $\rho$ and fall in $\tau_s$ are offsetting, and therefore their demand is unchanged.)

The properties of an "ideal tax" are now clear: such a tax would reduce demand without affecting the costs on which $(p_i/\psi)^D$ are based. Additional imposition of a tax of this sort could return the economy to the previous demand without additional

[1] Bent Hansen has called such steady-states "quasi-equilibria."

inflation. Generally, with such a tax, appropriate tax-*cum*-subsidy schemes can achieve *any* level of demand with *any* level of inflation. Economists have models in which the profits tax, the land tax, and the poll tax, accordingly, do not influence the cost-base of $(p_i/\psi)^D$. In contrast, in the last example, increases in the income tax added to $\omega^*$ and therefore indirectly to costs.

### D. *Example IV*

As a final example, consider a different industrial structure, in which firms and unions jointly set prices and wages to maximize the product of their utility functions (the Nash bargaining solution.) Let the utility of firms be proportional to profits. Let demand be as given in Example I, with production function $Q=L^{\epsilon_2}$, and with raw material requirements costing $\rho$, in real terms, per unit of output. Let utility of unions $V$ depend on wages and employment as $V=\omega^{\nu}L^{\epsilon_2}$. We find:

$$(11)\quad (p/\psi)^D = \frac{\rho}{1-\tau_s}\left(\frac{\epsilon_1+\epsilon_2}{1+\epsilon_1} - \frac{\nu}{1+\nu}\right) \cdot \left(\frac{\mu\epsilon_1 + (\mu-1)\epsilon_2}{(1+\epsilon_1)} - \frac{\nu}{1+\nu}\right)^{-1}.$$

As one result, note that a parallel calculation shows $(p/\psi)^D$ independent of the profits tax: the sales tax is more inflationary.

## III. Conclusion

This paper has given a method, illustrated by four examples, whereby inflation can be derived from microeconomic structure. This method is useful in determining how shifts in this structure cause shifts in inflation. Such examples are important, both to understand the assumptions made by econometricians, who implicitly use the Swedish method, and also to appreciate the range of possible conclusions to be derived from plausible microeconomic systems.

## REFERENCES

**B. Hansen**, *A Study in the Theory of Inflation*, London 1951.

**E. Lindahl**, *Studies in the Theory of Money and Capital*, London 1939.

**E. Kuh**, "A Productivity Theory of Wage Levels—An Alternative to the Phillips Curve," *Rev. Econ. Stud.*, Oct. 1967, *34*, 333–60.

# SAVINGS BEHAVIOR: NEW INFLUENCES AND CONSEQUENCES

# Social Security and Saving: The Extended Life Cycle Theory

By MARTIN FELDSTEIN*

The life cycle model is the central idea in the modern theory of saving because it provides the crucial link between the microeconomics of rational household behavior and the macroeconomics of the rate of saving. The fundamental insight of this theory, that aggregate saving is positive in a growing economy because the younger workers who save are more numerous and have higher earnings than the older retirees who dissave was presented by Sir Roy Harrod in the second lecture of his famous book, *Towards a Dynamic Economics* (1948). Harrod's description of the household's optimizing behavior, which he notes is an extension of Irving Fisher's (1930) analysis, is remarkably modern and "neoclassical" for someone who is rightly regarded as one of the great developers of Keynesian economic theory. But it was Franco Modigliani and his collaborators (e.g., 1954, 1957, 1963, and 1966) who developed Harrod's metaphor of "hump saving" into a quantitative theory and began the process of empirical verification that has made the life cycle model a central feature of our current economic understanding. Today I want to discuss an extension and verification of the life cycle model and its aggregate implications to include social security, the most important form of "saving" for the majority of American families.[1]

Social security tax "contributions" in the 1974–75 fiscal year were $83 billion, approximately equal to the $85 billion of personal saving and nearly 75 percent of total private saving (including corporate saving). I recently estimated that social security "wealth" (i.e., the actuarial present value of the future social security benefits of the current adult population) was more than $2 trillion in 1971 and thus nearly 60 percent of all other household assets (Feldstein 1974a). To ignore social security in the theory and testing of saving behavior is therefore to omit a factor of fundamental importance. Moreover, social security wealth is an exogenous variable that can in principle provide a better test of the life cycle hypothesis than is possible with the traditional measure of wealth.

## I. The Extended Life Cycle Model

The familiar life cycle model emphasizes the process of saving during working years for consumption during retirement. An important assumption of this model is that the period of retirement is fixed:

* Professor of economics, Harvard University, I am grateful to the National Science Foundation for its financial support.

[1] Theories and empirical tests have ignored the existence of corporate saving which now accounts for more than 40 percent of net private saving. I have examined the impact of corporate saving on total private saving (Feldstein 1973; Feldstein and Fane 1973) but will not discuss the results of those analyses in the current paper.

changes in wage rates, wealth or other variables may alter the allocation of lifetime consumption between working years and retirement years but they do not change the allocation of the years themselves between work and retirement.

The assumption of a fixed retirement period conflicts with observed behavior and has unintentionally caused some important evidence to be incorrectly interpreted as conflicting with the basic notion of rational life cycle saving. In 1900, only 37 percent of men over age 65 were retired; by 1971, this had doubled to 74 percent. It is appropriate therefore to extend the life cycle model to make the period of retirement endogenous.

A general formulation of such a model would have the individual choose both labor supply and consumption every year. A restricted specification, more in the spirit of the original life cycle model, would define a "preretirement period" during which the individual's labor supply is fixed and a "retirement period" during which the individual can vary his labor supply. The individual's preretirement consumption and saving and his "retirement period" labor supply would then be optimized together. In this extended life cycle model, the change in any endogenous variable has two separate effects on saving: first, it changes saving directly as it would in the traditional life cycle model, and second, by changing retirement, it alters saving indirectly.

Consider the example of social security. In the traditional life cycle model, the introduction of a public program that provides retirement benefits unambiguously reduces the amount of personal saving. However, in the extended life cycle model, there is a second effect of social security that tends to increase personal saving. By providing transfer payments to older persons who retire, social security induces the aged to reduce their supply of labor. This reduction in working years and the resulting increase in the period of retirement induce additional saving. The net effect of social security on the saving of the non-aged is indeterminate and depends on the relative strength of the traditional "saving replacement effect" and the new "induced retirement effect."[2]

There is a further reason, although one which I consider to be much less important, why the effect of social security is theoretically indeterminate. Robert Barro (1974), Levis Kochin (1974), and Merton Miller and Charles Upton (1974) recently presented an alternative extension of the life cycle model with very different implications about the effect of social security and of public debt. The novel feature of these models is the presence of planned bequests and the assumption that each individual's utility depends not only on his own consumption but also on the utility of all future generations. Although the introduction of social security creates wealth for the current generation, it also creates a liability (according to these authors) for a future generation or generations. The current generation, according to this theory, will resist this redistribution from their heirs to themselves by increasing their bequests by just enough so that their heirs are no worse off. The extra saving for this bequest just offsets the reduced saving that would otherwise be associated with the creation of social security benefits.

Although I think this model is ingenious, I find the analysis unconvincing in its argument that social security should have no effect on private saving.[3] First, the

[2] The extended life cycle model was presented in Feldstein (1974a). A formal proof that an increase in social security benefits may actually increase saving if the extent of retirement is endogenous is presented in Feldstein (1974b).

[3] I have developed the following criticism more fully in Feldstein (1975b).

theory is formally correct only if the social security program does impose a future liability exactly equal in present value to the wealth that is created for the current generation. Such a liability will be imposed only if the social security program will be abolished at some date in the future or the economy has no growth of either population or real per capita income. In the more relevant case of a growing economy and a social security program that will last indefinitely, each future generation can receive more in benefits than it pays in social security taxes. More specifically, in a steadily growing economy the individual earns on his social security taxes an implicit "rate of return" that is equal to the rate of growth of the tax base. If this rate of growth were equal to the marginal product of capital, every future generation would be as well off participating in social security as it would be otherwise; in this case there is no future liability at all.[4] In the more realistic case in which the rate of growth is less than the marginal product of capital, the current social security wealth entails a future liability with present value equal to some fraction of the current social security wealth. Thus, even with this extreme form of rationality, the social security program would represent net wealth and should have a depressing effect on saving if retirement is fixed.

This form of analysis ignores the effect of personal and corporate income taxes. The individual does not earn the marginal product of capital on his saving but a substantially lower rate of interest. The implicit "rate of return" on social security tax payments (i.e., the growth rate) may well equal or exceed the real after-tax rate of interest earned by most of the population. If the individual makes *this* comparison, he may well conclude that social security will provide a net gain for all of his future heirs. Instead of increasing his bequest, such an individual might actually decrease it. To avoid this mistake, the farsighted individual must recognize that the national reduction in real saving that follows the introduction of social security will lower the future capital stock which will in turn reduce future income tax collections and lower future real wage rates. In these indirect ways his heirs will be made worse off.

The complexity of these anticipations casts doubt on the empirical relevance of this entire exercise. To make the "correct" offsetting adjustment to his bequest, the typical individual must understand the effect of his bequest and of the social security program on all future generations and must assume that all of his future heirs will share his farsightedness. The individual must also understand the general equilibrium effect of reduced aggregate saving on future wages and taxes. Since the economics profession has previously ignored the need to adjust bequests to compensate for future tax liabilities, it may be safe to assume that the typical household has also not made the adjustment implied by the new theory.

Finally, it is clear that for most families bequests are very small and may even be unintentional. The typical voluntary "intergenerational transfers" are not bequests at death but the support of the consumption that their heirs enjoy as children. The nature of these transfers makes it even less likely that households will respond as the theory suggests. Moreover, if households do respond to an increase in social security by increasing such intergeneration transfers, this merely changes the nature of the individual consumption and does not constitute a transfer of real capital.

At most, therefore, the new emphasis

[4] This point was implicit in Paul Samuelson (1958) and developed explicitly by David Cass and Menachem Yaari (1967).

on rational bequests provides a further reason why the traditional life cycle model overstates the likely effect of social security on saving. The actual effect, i.e., the balancing of traditional wealth replacement, induced retirement, and increased bequests, is an impirical question. However, before turning to the econometric evidence, I must comment on the radical theory that Paul David and John Scadding (1974) recently presented.

The starting point for David and Scadding was the finding by Edward Denison (1958) that the gross private saving rate, i.e., the ratio of private saving plus depreciation to gross national product, was the same in 1929 and in the high employment years between 1948 and 1956. David and Scadding show that if Denison's ratio is modified by adding the purchase of consumer durables to gross private saving and by adding the gross imputed rental flow of consumer durables to *GNP*, the new ratio shows no trend during the period since 1900.

Although I think it is impossible to provide any direct behavioral model that relates gross saving to *GNP* and ignores all taxes, David and Scadding attempt to do just that. According to their theory, households treat all government expenditure financed by taxes as equivalent to private consumption and thus substitute tax-financed government expenditure for consumption with no change in saving; in contrast, households regard all government expenditures financed by government debt as equivalent to private investment and therefore allow each dollar of government debt to crowd out one dollar of private debt, i.e., private saving is again unchanged. In addition, David and Scadding offer no explanation of why households are concerned with *gross* saving instead of subtracting depreciation to obtain the real accumulation of wealth, a very important difference in establishing the stability of their saving ratio. Finally, by a stroke of poetic license, they describe this combination of implausible assumptions as a model of "ultrarationality."

While I believe that this specific theory of ultrarationality is unacceptable, the statistical evidence of a trendless saving ratio deserves comment. It is tempting to think that even if no acceptable theoretical explanation is available now, the stability of the saving ratio is too unlikely to be due to chance. But is it? The adjusted gross private saving ratio of David and Scadding, i.e., the ratio of private saving as normally defined plus depreciation plus consumer durable expenditures to the sum of gross national product and the implicit gross rental flow on consumer durables, is only one of a great many ratios that might be examined and it is certainly not the most natural of them. In a large enough set of alternatives, one can expect to find the most unlikely things. Moreover, as Peggy Musgrave (1975) has recently pointed out, the official national income account definition of private saving ignores the very substantial saving by the wholly owned foreign subsidiaries of *U.S.* corporations, now approximately 1.8 percent of *GNP*. Since this figure represents *net* saving, allowing for depreciation would make the effect even larger.

I believe that the general long-run stability of the particular ratio studied by David and Scadding is the result of a number of offsetting forces, including the growth of taxes, the higher level of incomes, the increase in retirement and the expansion of social security. There is no reason to expect this particular stability to continue in the future.

Let me turn now to the econometric analysis of the extended life cycle model. I will start not with the new research that has been done to implement this model but with the use of the model to

reinterpret some empirical results that have puzzled economists for more than a decade.

## II. Explaining the Paradoxical Effect of Pension Coverage on Direct Savings

The extended life cycle model can resolve the apparent paradox of the finding by George Katona (1965) and Phillip Cagan (1965) that persons covered by private pensions do not save less directly and may save more than those persons who are not covered by pensions.[5] Cagan explained his surprising results in terms of a "recognition effect": when an individual is forced to participate in a pension plan, he recognizes for the first time the importance of saving for his old age. Participation in a pension plan has an educational effect: more formally, it changes the individual's utility function as he perceives it *ex ante* during his working years. Katona added to this a second explanation: the "goal gradient" hypothesis borrowed from psychological research on the forming of aspirations. According to this theory, "effort is intensified the closer one is to one's goal" (Katona 1965, p. 4). In more conventional economic terms, this would imply that individual preferences are themselves a function of the opportunity set or of the initial position, a dramatic departure from the usual assumption of economic analysis.

The extended life cycle model explains the findings of Cagan and Katona without invoking a recognition effect or a model of changing preferences. Workers who are covered by pensions have an incentive to retire earlier than they otherwise would. The pension therefore has two effects on direct personal savings: (1) It reduces personal saving because it substitutes for household assets, but (2) it also increases personal saving because it lengthens the period of retirement over which accumulated assets will be spread. Since the net effect of the pension depends on the relative strength of these two forces, there is no paradox in the empirical finding that private pension coverage does not reduce direct personal saving.

Thus, the extended life cycle model incorporates the important structural feature of changing retirement and reconciles the previously paradoxical evidence on the effect of pensions with a theory of rational savings and retirement choice. The extended life cycle model also implies that social security has an ambiguous effect on private saving that can only be resolved by empirical analysis. Let me therefore turn to the recent econometric research on this subject.

## III. Recent Econometric Studies

In the first direct test of the extended life cycle model, I examined savings behavior in the United States from 1929 to 1971 (Feldstein 1974a). The analysis employed a generalization of the consumption function specification that Albert Ando and Franco Modigliani (1963) had used to test the traditional life cycle model. I reasoned that the effect of social security was most appropriately represented by the present actuarial value of the retirement and survivor benefits to which the current adult population was entitled; I called this magnitude social security wealth. The growth and extension of the social security program has raised social security wealth from 75 percent of *GNP* in 1940 to more than 200 percent of *GNP* today.

The social security wealth variable should play the same role in the aggregate consumption function that is expected of the ordinary "fungible wealth" variable:

[5] Alicia Munnell's (1974b) reanalysis of the Cagan data in a multiple regression model found a very weak and completely insignificant negative effect of pension coverage on direct saving.

a higher level of wealth should increase current consumption and decrease current saving. In addition to this direct effect, the growth of social security wealth should increase retirement and thus stimulate saving. The coefficient of the social security variable should therefore reflect the net effect of these two influences.

The statistical estimates indicate that social security does reduce private saving. The estimated marginal propensity to consume social security wealth was generally significantly positive and not significantly different from the coefficient of ordinary wealth.[6] The implied magnitude of the effect of social security on saving is therefore very large. The point estimate of the coefficient of social security wealth indicates that personal saving in 1971 was approximately halved by social security, implying a reduction in total private saving (including corporate saving) of 38 percent. When the sample was restricted to the period since 1947, the coefficients remained quite similar but the standard errors became so large that the effects of both ordinary wealth and social security wealth were insignificant. This evidence thus provides preliminary support for the conclusion that social security substantially depresses saving, but indicates the need for research with new bodies of data that can provide more precise estimates.

The time series data was analyzed in a number of interesting ways by Munnell (1974a, b). She tested the effect of retirement explicity by modifying the conmetric rate of decline in the labor force participation of older men from 1900 to sumption function with social security wealth to allow the marginal propensity to consume out of disposable income to vary with the labor force participation of men over 65. Although this provides a very imperfect measure of the expected future retirement of current workers, the interaction variable always had the expected sign. With this method of adjusting for the induced retirement effect of social security on saving, the coefficient of social security wealth was closer to an estimate of the pure wealth substitution effect; Munnell's coefficient of social security wealth had been nearly 50 percent greater than my estimate was in an equation that did not try to separate the effect of induced retirement. Munnell's decomposition also permits explicit estimates of the way in which the social security wealth replacement effect and the general increase in retirement have had offsetting effects on aggregate saving: in 1969, according to her estimates, social security wealth reduced personal saving by $54 billion while the greater retirement since 1929 increased saving by $26 billion.[7] In interpreting these figures it would of course be wrong to regard all of the impact of the increased retirement to be the indirect induced retirement effect of social security. Much of the increased retirement would

[6] The close similarity of the two coefficients reflects several offsetting effects and should not be regarded as evidence that social security wealth is an exact replacement for private wealth. Although social security wealth has an associated retirement effect that reduces its impact, it may be important that ordinary wealth may be of less certain value, may be intended in part for bequests, and is thus subject to estate tax. Ordinary wealth is also endogenous; a sustained period in which the unobservable disturbance in the consumption function is positive will reduce wealth and the resulting negative correlation between these two variables will bias the coefficient of ordinary wealth toward zero.

[7] This calculation is based on the first equation of Table 3, p. 562 of Munnell (1974a). It differs from Munnell's estimate which is based on her strange and extremely narrow concept of "retirement saving" which she defines to include only the increase in the net assets of life insurance companies and of private and government pension plans; by ignoring most forms of saving, Munnell greatly underestimates the saving effects of both social security wealth and changing retirement behavior. Her later work (Munnell 1975) uses only the more traditional definition of saving.

no doubt have occurred simply because of higher incomes, urbanization, the decline of self-employment, the depression, etc.; a simple extrapolation of the geo-1929 can account for nearly 75 percent of the increase in retirement from 1929 to 1969.[8]

A quite different type of evidence supporting the extended life cycle model is provided by an analysis of intercountry differences in saving rates. Modigliani (1970) has shown that the pattern of intercountry differences in private saving rates is consistent with the predictions of the traditional life cycle theory: higher saving rates in countries with higher rates of economic growth and higher proportions of the population of working age. To test the extended life cycle model, I (Feldstein 1974b) added measures of retirement behavior (the labor force participation rate of men over 65 and the life expectancy at age 65) and of the substitution effect of social security (the ratio of social security benefits per aged person to average income per capita). The coefficients of these variables had the predicted signs, were statistically significant in a variety of specifications, and accounted for a substantial portion of the variation in the saving rates of the 15 developed countries in the sample. In particular, the coefficient of the social security variable implied that the average level of social security benefits reduced the saving rate by 4.2 percentage points or one-third of the average private saving rate; similarly, an increase in relative social security benefits from one standard deviation below the mean to one standard deviation above reduced the private saving rate by 5.4 percentage points.

This of course reflects only the partial wealth replacement effect of social security since retirement is held constant statistically. However, the evidence indicates that the wealth replacement effect is much more important than the induced retirement effect. The next effect of social security implies that the average level of social security benefits reduces the saving rate by 3.5 percentage points, more than four-fifths of the pure wealth replacement effect.[9]

The extended life cycle model also resolves a long-standing puzzle in the international studies of saving behavior. A common finding of these studies has been that the countries with a higher per capita income also have a higher saving rate, an apparent conflict with the long-run homogeneity of the saving rate in the time series and with the presumption of the traditional life cycle theory that the division of lifetime consumption between working years and retirement years is independent of the overall level of consumption.[10] The estimates of the extended life cycle model show that the higher saving rate in higher income countries actually reflects a higher retirement rate. Average

[8] Although Munnell estimates a saving function rather than a consumption function, this is irrelevant since only the current value of income is included. She does, however, show that the results are not sensitive to the choice between the conventional national income accounting definition of saving and a more inclusive measure obtained as the annual change in Raymond Goldsmith's (1956) measure of household wealth; she updates the Goldsmith measure with data from the Securities and Exchange Commission.

[9] Henry Aaron (1967) reported a negative correlation between social security benefits and household saving rates but later (Aaron 1968) reported that this correlation could not be confirmed with slightly different data. His original study did not use a consistent life cycle framework and no specific estimates are reported in the 1968 study. Aaron's primary interest was the hypothesis that the saving rate influences the percentage of national income spent on social insurance; he concluded that the evidence does not support this hypothesis.

[10] This has been an implicit presumption but it reflects the further assumption of homothetic preferences, and it is not a fundamental implication of the life cycle model.

income has no effect on saving when labor force participation is held constant. In contrast, higher income induces more retirement, which in turn leads to a higher rate of saving.

The use of microeconomic household data to test the extended life cycle model and to assess the impact of social security is just beginning. There are now several data sets that could provide useful information. Munnell (1975) found strong evidence that men aged 45 to 65 substantially reduce their own saving if they are covered by social security or by a private pension.[11] Her analysis used an extended life cycle model that explicitly included the expected time to retirement and life expectancy after retirement, but there was no specific test of the effect of differences in expected retirement date. Because social security now covers almost everyone (the exceptions are almost all government employees or railroad workers with their own pension programs), the estimated effect of social security coverage is difficult to interpret. It is reassuring therefore that Munnell finds that saving is reduced by private pension coverage and varies inversely with crude estimates of pension benefits and social security benefits.

The effect of social security on private saving also explains why the concentration of wealth as traditionally measured has remained stable during the past fifty years in spite of strong economic pressures toward greater equality. Simon Kuznets (1956) calculated that the top 1 percent of the population received 15.6 percent of disposable income in the 1920's but only 7.7 percent in 1946 (the last year of his analysis). Although exactly comparable figures are not available for more recent years, there is no evidence of an increasing concentration of income and some evidence that the share received by upper income families has continued to decline. In contrast, Robert Lampman's (1962) classic study concluded that the share of wealth held by the top 2 percent of families varied only from 32 percent in 1922 to 29 percent in 1953; more recent evidence shows no decrease in concentration in the 1960's. It seems at first a paradox that the concentration of wealth has remained unchanged in spite of the reduced concentration of income and the rapid increase in estate and gift tax rates. The paradox is easily resolved however by recognizing that the vast majority of middle income and lower income households have substituted social security wealth for ordinary fungible wealth. I used the 1962 Federal Reserve Board *Survey of Consumer Finances* to compare the distribution of fungible wealth (i.e., excluding social security) with the distribution of total wealth including a detailed estimate of each family's social security wealth (Feldstein 1974c). The results show that the distribution of total wealth is much less concentrated than the distribution of fungible wealth and has therefore become much more equal during the past half century. For example, the top one percent of families with a head between 35 and 64 years old owned 28.4 percent of fungible wealth but only 18.9 percent of total wealth; similarly, the top 4 percent of such families had 45 percent of fungible wealth but only 31 percent of total wealth.

This analysis of the distribution of social security wealth also provides indirect support for the traditional life cycle model. Within each age group, the distribution of income among income classes was more similar to the distribution of

[11] Her analysis is based on the National Longitudinal Survey directed by Herbert Parnes and conducted by the Bureau of the Census. Saving is defined as the change in net worth over a 3 year period.

total wealth than to the distribution of fungible wealth.[12]

## IV. Conclusion

With our current pay-as-you-go method of financing social security, the social security tax receipts are paid out as concurrent benefits and are not accumulated. When we, the current generation of workers, retire, we will not receive social security benefits by drawing on an accumulated fund. Instead, our benefits will depend on the tax payments of those who are at work when we retire. Because there is no accumulated social security fund, the decrease in private saving that is caused by social security entails an equal decrease in national saving.

The very substantial size of the estimated decrease in saving implies a correspondingly large decrease in the nation's capital stock. The parameter estimates in my study of *U.S.* time series data implied that without social security the *U.S.* capital stock would eventually be some 60 percent higher (Feldstein 1974a). The larger capital stock would mean greater productivity, higher real wage rates and a higher national income.[13] But the important implication of the reduction in saving is not the fall in income or wages per se. The reduction in welfare comes from the distortion in saving, the substitution of an asset with a very low implicit rate of return for real capital accumulation with a much higher social rate of return.[14]

[12] Although the focus of this survey is on savings, I should mention some of the recent related work that has shown that social security has a powerful effect on retirement behavior. Building on earlier work by Margaret Gordon (1963) and Henry Aaron (1967, 1968), I used a cross section of countries to estimate the effect of social security on retirement (Feldstein 1974b). Much more detailed analyses of microeconomic data for the United States have been presented by Joseph Quinn (1975) and Michael Boskin (1975).

[13] This assumes, I believe correctly, that the *U.S.* economy could absorb a higher rate of capital accumulation. The early Keynesian "secular stagnationists" saw social security as a welcome method of reducing the excess saving that they feared would cause a permanent depression after World War II; see Seymour Harris (1941).

[14] For a more general discussion of these issues, see Feldstein (1975a).

## REFERENCES

**H. J. Aaron,** "Social Security: International Comparison" in O. Eckstein (ed.), *Studies in the Economics of Income Maintenance,* Washington 1967.

———, "International Comparisons," Appendix D in J. Pechman, H. Aaron, and M. Taussig, *Social Security: Perspectives for Reform,* Washington 1968.

**A. Ando and F. Modigliani,** "The 'Life Cycle' Hypothesis of Saving: Aggregate Implications and Tests," *Amer. Econ. Rev.,* Mar. 1963, *53,* 55–84.

**R. Barro,** "Are Government Bonds Net Wealth?" *J. Polit. Econ.,* Nov./Dec. 1974, *82,* no. 6, 1095–1117.

**M. J. Boskin,** "Social Security and Retirement Decisions," mimeo 1975.

**P. Cagan,** *The Effect of Pension Plans as Aggregate Savings,* New York 1965.

**D. Cass and M. F. Yaari,** "Individual Saving, Aggregate Capital Accumulation, and Efficient Growth," in K. Shell (ed.), *Essays in the Theory of Optimal Growth,* Cambridge 1967.

**P. A. David and J. L. Scadding,** "Private Savings: Ultrarationality, Aggregation and 'Denison's Law,'" *J. Polit. Econ.,* Mar./Apr. 1974, *82,* no. 2, part I.

**E. F. Denison,** "A Note on Private Saving," *Rev. Econ. Statist.,* Aug. 1958, *40,* 261–67.

**M. S. Feldstein,** "Tax Incentives, Corporate Saving, and Capital Accumulation in the United States," *J. Publ. Econ.,* 1973, *2,* 159–171.

———, "Social Security, Induced Retirement and Aggregate Capital Accumulation," *J. Polit. Econ.,* Sept./Oct. 1974a, *82,* no. 5.

———, "Social Security and Private Savings: International Evidence in an Extended Life Cycle Model," (1974b) in *The Eco-*

*nomics of Public Services*, an International Economic Association Conference Volume, M. Feldstein and R. Inman (eds.), forthcoming.

———, "Social Security and the Distribution of Wealth," Harvard Institute of Economic Research disc. pap. no. 370, June 1974c.

———, "Toward a Reform of Social Security," *The Public Interest*, Summer 1975a, no. 40.

———, "Perceived Wealth in Bonds and Social Security: A Comment," *J. Polit. Econ.*, forthcoming, 1975b.

——— and G. Fane, "Taxes, Corporate Dividend Policy and Personal Savings: The British Postwar Experience," *Rev. Econ. Statist.*, Nov. 1973, *55*, 399–411.

I. Fisher, *The Theory of Interest*, New York 1930.

R. W. Goldsmith, *A Study of Saving in the United States*, Princeton 1956.

M. S. Gordon, "Income Security Programs and the Propensity to Retire," University of California Institute of Industrial Relations 1973, reprint no. 210.

S. Harris, *Economics of Social Security*, New York 1941.

R. F. Harrod, *Towards a Dynamic Economics*, London 1948.

G. Katona, *Private Pensions and Individual Saving*, Ann Arbor 1965.

L. Kochin, "Are Future Taxes Anticipated by Consumers?" *J. of Money, Credit and Banking*, Aug. 1974, *6*, 385–94.

S. Kuznets, *Shares of Upper Income Groups in Income and Savings*, New York 1956.

R. Lampman, *The Share of Top Wealth-Holders in National Wealth, 1922–56*, Princeton 1962.

M. H. Miller and C. W. Upton, *Macroeconomics: A Neo-classical Introduction*, Homewood 1974.

F. Modigliani, "The Life Cycle Hypothesis of Saving, the Demand for Wealth and the Supply of Capital," *Social Research*, Summer 1966, *33*, no. 2, 160–217.

———, "The Life Cycle Hypothesis of Saving and Intercountry Differences in the Saving Ratio" in W. A. Eltis et al. (ed.), *Induction, Growth and Trade, Essays in Honour of Sir Roy Harrod*, Oxford 1970.

——— and A. Ando, "Tests of the Life Cycle Hypothesis of Savings," *Bulletin of the Oxford University Institute of Statistics*, May 1957, *19*, 99–124.

——— and R. Brumberg, "Utility Analysis and the Consumption Function: An Interpretation of Cross-Section Data" in K. Kurihara (ed.), *Post Keynesian Economics*, New Brunswick 1954.

A. H. Munnell, "The Impact of Social Security on Personal Savings," *National Tax Journal*, Dec. 1974a, *27*, no. 4.

———, *The Effect of Social Security on Personal Savings*, Cambridge 1974b.

———, "Private Pensions and Saving: New Evidence," *J. Polit. Econ.*, forthcoming.

P. B. Musgrave, "Foreign Investment in the National Income Accounts," mimeo., Oct. 1975.

Joseph F. Quinn, "The Microeconomics of Early Retirement: a Cross-Sectional View," Sept. 1975, Boston College Dept. of Economics Working Papers in Economics, no. 72.

P. A. Samuelson, "An Exact Consumption-Loan Model of Interest with or without the Social Contrivance of Money," *J. Polit. Econ.*, Dec. 1958, *66*, no. 6, 467–82.

# Intergenerational Transfers and Life Cycle Consumption

*By* ALAN S. BLINDER*

It is now roughly twenty years since Franco Modigliani and Milton Friedman revived and gave empirical substance to Irving Fisher's theory of lifetime planning by the consumer.[1] This paper can be thought of as a discussion of what it means to take the life-cycle model seriously. In this spirit, Section II explores the role of intergenerational transfers and, in particular, the wealth elasticity of bequests in the aggregate consumption function. However, it would be a mistake to take the life cycle hypothesis *too seriously*. Its central message—that the lifetime pattern of consumption is independent of the lifetime pattern of earnings—cannot be literally true in a world where there is uncertainty, where working and consuming are competing uses of time, and where the ability to borrow against future earnings is severely limited.[2] Section I concentrates on capital market constraints, stressing how intergenerational transfers help determine which households are constrained and which are not. Though the orientation is clearly theoretical, the relevance to public policy cannot be missed in view of recent fiscal policy actions.

## I. Consumption Behavior of Planners and Reactors

While there are many principles according to which households can be divided, I find it useful to think of the aggregate consumption function as a weighted average of the consumption functions of two types: planners and reactors.

Planners have sufficient financial means to adhere tolerably close to their optimal consumption paths. Because borrowing constraints are not binding for them, the relevant budget constraint is based on "wealth" or "life resources," defined as the sum of discounted earnings and financial inheritances. It is easy to show that the lifetime marginal propensity to consume (*MPC*) out of wealth is positive, and is smaller than unity if there is a bequest motive.[3] A planner will spread the extra consumption permitted by any windfall gain, such as a government rebate check, over all the remaining years of his life, and probably will add some of it to his bequest. Since the short-run *MPC* depends critically on age, one good way to elicit a substantial spending response from a one-time transfer payment is to concentrate the funds on elderly households. In this light, the recent $50 per head social security bonus looks like a highly imaginative fiscal policy.

Reactors are consumer units that are prevented from pursuing their first-best optimal plans by borrowing constraints. A household which would like to spend more than its income, but is prevented from so

* Assistant Professor of Economics, Princeton University. I have benefited from helpful comments of Martin Hellwig, William Buiter and Sherman Robinson, who are not to be implicated in the conclusions. Financial support from the National Science Foundation is gratefully acknowledged.

[1] No distinction will be made between the life-cycle and permanent income models, as they are just two ways of implementing the Fisherian analysis empirically.

[2] Regarding uncertainty, see Keizo Nagatani. Regarding time budgets, see James Heckman. Regarding borrowing constraints see Tsuneo Ishikawa (1974), Thomas Russell, James Tobin and Walter Dolde, John Flemming or Clark Wiseman.

[3] See the author (1974), pp. 39–40.

doing by imperfect capital markets, will generally consume its entire income. A windfall loss, therefore, must come entirely out of consumption. A windfall gain, on the other hand, loosens the capital market constraint; so the short-run *MPC* may be high but probably will not be unity. Thus, the reactor's responses to windfall gains and losses are asymmetric.

### *A. Who Plans and Who Reacts?*

While it is tempting to think of the poor as reactors and the rich as planners, such a simple identification is not generally correct. As Tsuneo Ishikawa (1974) showed, the classification has a lot to do with the timing and volume of intergenerational transfers.[4]

The poor man with a flat age-income profile and no financial inheritance is constrained by many things but not by the inability to borrow against future receipts. The households who feel the pinch of capital market imperfections most acutely are those anticipating large cash inflows later in life—whether from high earnings or from large financial inheritances—but whose current cash inflows are insufficient to finance the consumption stream warranted by their lifetime resources.

Wealthy households which expect a substantial financial inheritance late in life when their parents die will probably be reactors for much of their lives. Then, when the inheritance is finally received, diminishing returns will dissuade them from completely making up for lost time. So capital-market imperfections will induce them to consume less and save more over their lifetimes. Of course, parents can ease these constraints by transferring wealth *inter vivos*, but this would simply shift the burden of capital-market constraints onto the parent generation as it sought to accumulate wealth at a much earlier age.

Intergenerational transfers of human capital are no doubt much more important quantitatively than are financial inheritances. Further, they have the important characteristic that children must receive them at relatively young ages if they are to do any good. Child care expenses cannot be deferred until the children are grown. College education yields small financial benefits if taken at age fifty. Since intergenerational transfers of human capital must be made rather early in life, they tend to subject the donor generation to capital market constraints. Yet, from the point of view of the recipients, human capital inheritances are effectively received late in life because they manifest themselves in steeply-rising age-earnings profiles. Thus the recipients become subject to borrowing limitations as well. By contrast, persons less well-endowed with human capital typically have flat age-earnings profiles. In this respect, then, the upper (labor) income groups are more likely to be reactors than planners. At least where young families are concerned, redistribution from poor to rich may well increase total spending!

### *B. Social Security as an Example*

The Social Security System is probably the biggest program of intergenerational financial transfers in our society, and it illustrates the usefulness of both the life-cycle perspective and the planner-reactor dichotomy. Without taking the life-cycle point of view, one might expect such a program of forced saving to reduce voluntary private saving roughly dollar for dollar. Yet the data seem to refute this.

Martin Feldstein and Alicia Munnell used the life-cycle model to provide one plausible explanation for this "paradox": if the social security system induces earlier retirement, more intra-life-cycle savings will be needed to finance the longer retirement periods. Here is another, applicable

[4] There are other determinants of which households are constrained, such as unanticipated spells of unemployment, but these are beyond the purview of this paper.

in the case of planners. Increases in social security benefits are generally not fully funded. Planners thus correctly perceive increases in social security as *negative bequests*, that is, as forced transfers from their children to them. As Gary Becker and Robert Barro have pointed out, the rational response is to increase their previously planned bequests so as to compensate their children for the government's action. To do this, they will save more.

If social security induces more workers to retire abruptly at age 65, as Feldstein and Munnell suggest, then optimal consumption theory provides still another explanation for the paradox. Since leisure time and consumption are presumably complementary, the abrupt upward movement in leisure time at retirement will make a corresponding abrupt upward movement in consumer spending optimal. This means "tilting" the lifetime consumption path towards lower consumption, i.e., higher savings, in the preretirement years.[5]

For reactors, there is an entirely different reason to expect increased levels of social security benefits and taxes to induce greater private saving. Nonfunded social security benefits raise life resources and thus desired consumption, while higher payroll taxes reduce current take-home pay. This forces households constrained by borrowing limits to curtail consumption. Note also that by transforming some households from planners to reactors, social security increases the income-sensitivity of the aggregate consumption function.

## II. The Wealth Elasticity of Bequests

As the size and timing of bequests are such crucial determinants of life cycle saving behavior, it is worth pausing to inquire into the motives for bequests, the form of bequests, and how the optimal bequest varies with wealth.

### *A. Motives for Bequests*

The conventional approach is to consider the bequest—or, what amounts to approximately the same thing, the wealth or utility of the beneficiaries—as another argument of the utility function. This hypothesis has a certain intuitive appeal, and may be descriptively accurate Equally important, economists have been developing the utility-maximization model for 100 years and by now can do quite a lot with it.

Becker's model of intergenerational transfers is a case in point. By some reasonable-sounding assumptions, Becker is able to demonstrate that the wealth elasticity of bequests must exceed unity—an empirical proposition which, if true, has some very fundamental consequences (see below). His argument can be paraphrased as follows. Suppose that the current family head acts *as if* he were maximizing a homothetic utility function with the wealths of the current and all future generations as arguments, and that capital markets are sufficiently perfect to permit him to do this. Then he will want to spread any increment to his own wealth proportionately over all generations. This means that he will consume only a small fraction of any marginal addition to wealth, bequeathing the rest. Since bequests are a small fraction of wealth, the *percentage* rise in bequests will be very large.

Becker's assumptions are less innocuous than they seem. Suppose instead that people only allocate *their own* life resources and that the utility function is not necessarily homothetic. The following function is sufficiently general to illustrate the point:

$$U = \frac{C^{1-\delta}}{1-\delta} + \phi\frac{B^{1-\beta}}{1-\beta},$$

where $C$ is own consumption and $B$ is the

[5] If $U(C,L)$ is the utility function, where $C$ is consumption and $L$ is leisure time, then $U_C(C,L)$ is a continuous function of time along an optimal path. Assuming $U_{CL}(\cdot) > 0$ and $U_{CC}(\cdot) < 0$, a discontinuous upward increment in $L$ necessitates a discontinuous upward movement in $C$.

bequest. Indifference curves are homothetic if and only if $\delta=\beta$, and in this case the wealth elasticity of bequests is unity. If, $\delta>\beta$ instead, the wealth elasticity of bequests exceeds one, and if $\delta<\beta$, the wealth elasticity is less than one.[6]

Hal Varian has suggested a different motive for bequests: that parents bequeath wealth to ensure that their offspring do not wind up too low in the income distribution. There is no obvious implication about the wealth elasticity of bequests, but the notion does imply that the average level of bequests increases with increasing inequality.

No mention has been made thus far of the number of children. Both of the aforementioned motives, and most others, suggest that the taste for bequests—the parameter $\phi$ in the utility function—would rise with the number of children. If this is the case, a slower population growth rate will lead to lower levels of bequests. However, with fewer human beings in which to invest, the composition of intergenerational transfers should shift more from human form to financial form, so financial bequests may not fall.

*B. Bequests in Human or Financial Form?*

Whether human or financial wealth is the medium for intergenerational transfers is important for macroeconomic reasons largely because bequests in human form usually entail either consumer expenditures in the national income accounting (*NIA*) sense, or withdrawals from the labor force, whereas financial bequests constitute *NIA* savings. Furthermore, discussion of the form in which bequests are left sheds insight on the wealth elasticity. This is because a compelling theoretical argument can be made that marginal adjustments in lifetime savings (i.e., bequests) will be made preponderantly in financial form by the wealthy and in human form by the not-so-wealthy.

The argument, which is developed formally by T. Ishikawa (1975) and Donald Parsons, is simple and intuitively appealing. Due to the finite capacity of human beings, bequests in human form are subject to the law of diminishing returns in that successive increments of financial sacrifice by the parents lead to smaller and smaller financial gains for the children. Bequests in financial form, by definition, cannot be subject to such diminishing returns.[7] Optimal behavior, then, will lead decedents to make all bequests in human form until the rate of return on human capital is reduced to the opportunity cost of funds, and thereafter to make all bequests in financial form. This is illustrated in Figure 1, Panel (a), where H* is the critical level of human wealth.

Apparently, most families do not reach H*, for such data as exist show that very few bequests are made in financial form. In a 1959 survey, 73 per cent of all household heads aged 55–64 reported receiving no inheritance, and only about 2 per cent reported receiving $25,000 or more. The average reported financial inheritance, $2,800 per household, must have been dwarfed by the typical human inheritance.[8]

Suppose we had data on bequests, and their division between human and nonhuman form, for a cross-section of families. What would we see? Since, the point H* varies among families because of differing discount rates, genetic endowments of human capital, numbers of children, and other things, we might observe smoothly concave and convex functions for human and financial bequests respectively, as shown in Panel (b). In a word, the elasticity of financial bequests with respect to

[6] For a proof, see the author (1974), pp. 35–37.

[7] Both, of course, may be subject to the law of diminishing marginal utility.

[8] I made these computations from the original data tape collected for the study by James Morgan et al.

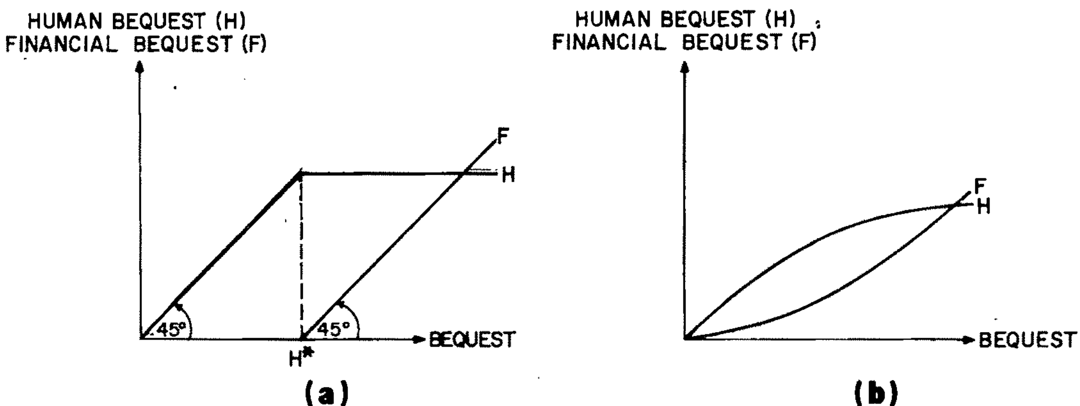



FIGURE 1

total bequests would exceed unity, so the wealth elasticity of financial bequests would exceed the wealth elasticity of total bequests. This is an important point since all efforts to measure the wealth elasticity of bequests (see below) have focused on financial bequests.

Perhaps offsetting this tendency in the upper income strata is the differential tax treatment of human and financial bequests. Intergenerational transfers of human capital are generally not taxed, and often are subsidized. By contrast, financial bequests above $60,000 are subject to a progressive Federal estate tax, and most states levy progressive inheritance taxes as well. Because of this progressivity, post-tax estates could be a concave function of lifetime wealth even if pre-tax bequests are a convex function. A descriptive regression fitted to grouped data on 1969 Federal estate tax returns is:

$$\log(\text{estate tax}) = -5.226 + 1.259 \log(\text{economic estate} - \$60{,}000)$$
$$(.308) \quad (.025)$$
$$R^2 = .995, \qquad \text{standard error} = .222,$$

which indicates substantial progressivity in marginal, as well as average, tax rates. However, this ignores the fact that accrued capital gains bequeathed at death escape taxation entirely.

Finally, note that the conclusion holds only for cross-section comparisons, not for time series trends. Schooling is a labor-intensive activity in which productivity gains are likely to be minimal. So the cost in dollars of purchasing the number of years of schooling represented by point H* in Figure 1 rises roughly at the same rate as per capita income.

### *C. Guesstimates of the Wealth Elasticity of Bequests*

Whether the wealth elasticity of bequests exceeds or falls short of unity governs whether the lifetime *MPC* falls or rises as we move up the income scale. This, in turn, has a number of important macroeconomic implications. Suppose, for concreteness, that the wealth elasticity of *financial* bequests is greater than one. Then a permanent equalization of the income distribution will permanently raise aggregate consumption. Also, since the savings rate will rise as per-capita wealth increases, a growing economy will have a rising capital-output ratio. If the aggregate elasticity of substitution is less than one, labor's share in national income will be steadily increasing. Each of these tendencies is reversed if the wealth elasticity of bequests turns out to be less than unity.

Of course, these tendencies are derived from the behavior of *inter*-life-cycle sav-

ings, which are likely to be dwarfed by *intra*-life-cycle savings if the population is growing rapidly. But if the brave new world of the future does indeed have zero population growth, these trends, which have been masked so long, will begin to come to the fore.

To date, lack of either time series or cross-section data on lifetime income and bequests has precluded direct measurement of the wealth elasticity of bequests. However, several indirect—and highly questionable—attempts have been made to estimate it.

One indirect way is to look for systematic differences in marginal propensities to consume *permanent income* by income class, as I did in my 1975 paper. I found that aggregate time series data for the United States since World War II have little power to discriminate among competing hypotheses. Such evidence as I did turn up consistently pointed to a wealth elasticity of bequests below unity; but I am not convinced. Real-world shifts in the income distribution confound transfers between rich and poor with transfers between young and old, male and female, black and white. Further, for reasons cited above, the rich may be reactors more frequently than planners. If so, their high *short-run MPC*s could be incorrectly labelled as high *long-run MPC*s, given the imperfect nature of the distributed-lag proxy for permanent income.

Using cross-section data by state, Hal Varian regresses the average bequest in the state ($b_i$) on average lifetime income in the state ($y_i$), income squared, and the interaction of income with a measure of income inequality ($P_i$, for the Pareto measure). This seemingly direct measurement also constitutes an indirect test because, except for the inequality measure, *none* of the data needed for Varian's study exist! His technique might well be called "regression without data." Varian uses *annual* income in the state as a proxy for *lifetime* income, and fabricates data for $b_i$ by the clever, but probably dubious, assumption that the average level of bequests among estates below $60,000 is the same in each state. He finds a negative coefficient on $y^2$, suggesting a concave $b(y)$ function, which is consistent with my findings. However, the negative coefficient on $P \cdot y$, which he interprets as support for his hypothesis,[9] would also be interpreted as evidence for a convex $b(y)$ function.

Finally, in a current Ph.D. dissertation, James Adams tries to estimate the wealth elasticity of bequests in a cross-section of individuals by using current net worth as a proxy for bequests and current earnings as a proxy for lifetime income. Given that both lefthand and righthand variables are proxies, it is hard to interpret his finding of an elasticity well in excess of unity. For example, it may just reflect the high elasticity of asset-holding with respect to earnings *within* the life cycle, despite his attempt to devise a statistical control for age.

### III. Concluding Remark

It is typically the case in economics that theory is needed to unearth the interesting questions, while empirical research is needed to answer them. So it is here. Despite some crude attempts using aggregate time series, a cross-section of states, and a cross-section of individuals, it must be admitted that we know relatively little about the wealth elasticity of bequests. In the case of capital market imperfections, the theoretical notions have yet to be given empirical substance. Twenty years after the seminal contributions of Modigliani and Friedman, the aggregate consumption function retains some of its mystery.

[9] The Pareto coefficient declines as inequality rises.

## REFERENCES

J. D. Adams, Ph.D. dissertation in preparation, University of Chicago.

R. J. **Barro,** "Are Government Bonds Net Wealth?," *J. Polit. Econ.*, Nov./Dec. 1974, *82*, 1095–1117.

G. S. **Becker,** "A Theory of Social Interactions," *J. Polit. Econ.*, Nov./Dec. 1974, *82*, 1063–94.

A. S. **Blinder,** *Toward an Economic Theory of Income Distribution*, Cambridge, 1974.

———, "Distribution Effects and the Aggregate Consumption Function," *J. Polit. Econ.*, June 1975, *83*, 447–76.

M. S. **Feldstein,** "Social Security, Induced Retirement, and Aggregate Capital Accumulation," *J. Polit. Econ.*, Sept./Oct. 1974, *82*, 905–26.

J. S. **Flemming,** "The Consumption Function when Capital Markets are Imperfect: The Permanent Income Hypothesis Reconsidered," *Oxford Econ. Pap.*, July 1973, *25*, 160–72.

M. **Friedman,** *A Theory of the Consumption Function*, Princeton 1957.

J. **Heckman,** "Life Cycle Consumption and Labor Supply: An Explanation of the Relationship Between Income and Consumption," *Amer. Econ. Rev.*, Mar. 1974, *64*, 188–94.

T. **Ishikawa,** "Imperfection in the Capital Market and the Institutional Arrangement of Inheritance," *Rev. Econ. Stud.*, July 1974, *41*, 383–404.

———, "Family Structures and Family Values in the Theory of Income Distribution," *J. Polit. Econ.*, Oct. 1975, *83*, 987–1008.

F. **Modigliani** and R. **Brumberg,** "Utility Analysis and the Consumption Function: An Interpretation of Cross-Section Data," in K. K. Kurihara (ed.), *Post-Keynesian Economics*, New Brunswick 1954.

J. N. **Morgan,** M. H. **David,** W. J. **Cohen** and H. **Brazer,** *Income and Welfare in the United States*, New York 1962.

A. H. **Munnell,** "The Impact of Social Security on Personal Savings," *Nat. Tax Jour.*, Dec. 1974, *27*, 553–67.

K. **Nagatani,** "Life Cycle Saving: Theory and Fact," *Amer. Econ. Rev.*, June 1972, *62*, 344–53.

D. O. **Parsons,** "A Family Model of Schooling," unpublished, Ohio State Univ., undated.

T. **Russell,** "The Effects of Improvements in the Consumer Loan Market," *J. Econ. Theory*, Nov. 1974, *9*, 327–39.

J. **Tobin** and W. **Dolde,** "Wealth, Liquidity and Consumption," in Federal Reserve Bank of Boston, *Consumer Spending and Monetary Policy: The Linkages*, Boston 1971.

H. R. **Varian,** "Inequality and the Motive for Bequests," unpublished, Massachusetts Institute of Technology, Nov. 1974.

C. **Wiseman,** "Windfalls and Consumption Under a Borrowing Constraint," *Rev. Econ. Statist.*, May 1975, *57*, 181–84.

# The Optimal Taxation of Commodities and Income

*By* DAVID F. BRADFORD AND HARVEY S. ROSEN*

The last few years have seen a resurgence of interest in the old question of how best to raise tax revenue. Roughly speaking, two different problems have been studied. The first is to find a set of commodity taxes that is optimal given certain efficiency and (sometimes) equity considerations. In a second strain of the literature, it is assumed that the revenue system is based upon income rather than commodity taxation, and the problem is to determine the optimal degree of progressivity (or regressivity).[1,2]

The principal motivation of some writers in the optimal taxation literature seems to be the discovery of fairly simple rules which policy makers actually can implement. Others are more interested in theoretical exploration of the implications of alternative economic assumptions than in developing usable policy recommendations. Practically all the contributions, however, have been quite mathematical and thus inaccessible to many practitioners in the public finance area. The purpose of this essay is to discuss in a nontechnical way the methodology and principal conclusions of the optimal taxation literature.[3]

In Sections I and II are discussed the optimal commodity and income tax literatures, respectively. Following this are some observations on the accomplishments of optimal taxation research and on some open questions.

## I. Optimal Commodity Taxation

Since the literature contains many and varied derivations of the principal theorems of optimal commodity taxation,[4] we shall not carry out detailed proofs here. We can point out, however, some of the variations in the way the problem is posed. Most commonly a revenue constraint is taken as a starting point, together with an assumption that the government must use per unit commodity taxes. Thus lump sum taxes are excluded. If $x_i$ is the quantity of the $i$th good purchased by the household sector from the production sector (negative if the households are net sellers, as in the case of the commodity "leisure"), and $T_i$ is the per unit tax, the revenue constraint is

$$(1) \qquad \sum T_i x_i = R,$$

where $R$ is the required revenue level.

The taxes are the difference between the

* U.S. Treasury Dept. and Princeton University, respectively. The authors would like to thank Roger Gordon for useful conversations and Jay Stuart for assistance in gathering material. An extended version of this paper is available upon request to the authors.

[1] There is some overlapping of these strains. For example, A. B. Atkinson and Joseph E. Stiglitz consider the problem of differential commodity taxation in the presence of an income tax.

[2] Although we shall focus upon these problems in this paper, the optimal tax literature has had a somewhat wider scope. For example, Peter A. Diamond and James A. Mirrlees consider the problem of optimal expenditure along with taxation.

[3] Consult Nicholas Stern or Atkinson and Stiglitz for more technical surveys.

[4] See, for example, Frank A. Ramsey, M. Boiteux, or Agnar Sandmo.

prices, $p_i$, received by producers and $P_i$, paid by the consumers,

$$T_i = P_i - p_i. \tag{2}$$

It is frequently assumed that producer prices are fixed, so that by setting taxes we set consumer prices and hence consumer welfare. The problem is then to make the choice of taxes in such a way as to maximize the resulting consumer welfare, i.e., to minimize excess burden.

A typical approach is to assume there to be only one consumer (hence no distribution problem), with a utility function $U(\cdot)$ depending on consumption vector $x$. Thus the objective might be to choose $P$ (a vector of consumer prices) and $p$ (a vector of producer prices) to

$$\text{Maximize } U(x(P)) \tag{3}$$

subject to

$$\sum_i x_i(P)P_i = 0^5$$

and to

$$U_i(x(P)) = aP_i. \tag{4}$$

Conditions (4) are the familiar first order implications of the household's optimization, with $a$ being the Lagrangian multiplier.

More often an indirect utility function $V(P, 0) = U(x(P))$ is used, (the zero argument draws attention to the assumption of no transfer income) because the derivations become very simple when use is made of "Roy's Identity"

$$\frac{\partial V}{\partial P_i} = -x_i(P)\cdot\frac{\partial V}{\partial M}, \tag{5}$$

where $\partial V/\partial M$ is the "income" derivative.

Putting these pieces together in any of several sequences leads to the famous Ramsey result on optimal commodity taxation:

$$\sum_i T_i S_{ik} = bx_k, \qquad k = 1, \ldots, m \tag{6}$$

where $S_{ik}$ is the derivative of the demand for the $i$th good with respect to the $k$th price, other prices and utility being held constant, and $b$ is independent of $k$. The lefthand side gives an estimate of the change in demand for the $k$th good which would occur if the taxes were removed. Hence (6) says that the proportional change in demand (thus estimated) should be the same for all commodities—the Ramsey result.

Conditions (6) can also be expressed in terms of elasticities. Probably the most familiar "optimal tax" result is the form which applies when the off-diagonal elasticities are zero. In this case the first order conditions associated with (3) lead to the "inverse elasticity rule":

$$t_r = \frac{d}{E_{rr}}, \qquad r = 1, \ldots, m. \tag{7}$$

where $t_r = T_r/P_r$, the percentage or ad valorem rate of tax, $d$ is a constant, and $E_{rr}$ is the elasticity of the ordinary (uncompensated) demand function for the $k$th good. This formula has certainly been of importance in forming economists' intuitions on tax and price regulatory questions. It underlies the notion of charging according to "what the traffic can bear" in transportation, for example, and is the basis for the acceptance on efficiency grounds of such taxes as those on tobacco and alcohol, the demand for which is presumed price inelastic.

An important application of the analysis is to the presumptive case for direct over indirect taxation. The classic Hotelling argument for marginal cost pricing seemed to some to support the conclusion that an "income tax" will involve no efficiency

[5] Paul A. Samuelson uses a somewhat different formulation of the problem and one which has the virtue of emphasizing the resource releasing function of the taxes.

cost. When it was recognized, however, that the "income" of the tax system is not the "budget level" of the elementary theory of consumer demand, but rather the product of a certain price, the wage, and a demanded quantity (negative net purchase of leisure), the apparent a priori advantage of an income tax was lost. The analyses of W. J. Corlett and D. C. Hague, I. M. D. Little and Milton Friedman to this effect all are applications of the theory of optimal commodity taxation as is neatly shown by Sandmo.

While the extensive subsequent work has shown how difficult it is to sustain *any* simple rules for commodity taxation, the result of the spreading awareness of this work has been to make economists think about tax questions in a new way and to hasten the search for rules which are reasonably robust.

For example, as Stiglitz and Atkinson point out, optimal tax analysis makes it clear that there is no a priori assurance that the income tax is the single best instrument for income redistribution—such "commodity taxes" as are represented by housing subsidies or food stamps might contribute to an optimal program. Michael Boskin notes that, in view of the differences in the observed elasticities of household supply of the two types of labor (husband labor and wife labor), it is probably efficient to tax these "commodities at different rates. Martin Feldstein (1975) uses the same basic approach to examine the choice between "tax expenditures" and direct expenditure methods of achieving an increase in a specified activity.

A natural question in view of the interpretation of the income tax as a commodity tax is whether taxation of labor only (i.e., uniform taxation of commodities) is appropriate. Not surprisingly, the answer is that it will be so when labor is inelastically supplied. Sandmo shows that this in turn will follow if utility is separable between leisure and all other goods and homogenous in those goods. Intuitively this separability means that further efficiency cannot be gained by differential taxation of goods that are "related" to leisure. Several writers have noted an important consequence when this result is reinterpreted in an intertemporal context. If utility is a function of consumption and leisure at different dates and separability obtains, then no taxes on interest income should be levied—consumption is the appropriate tax base. This simply illustrates the challenge, implicit in the optimal tax approach, to the widespread acceptance of taxation on the basis of Haig-Simons income which has been emphasized by Feldstein (forthcoming).

While an "income tax" can be regarded as a tax on the sale of labor (negative net purchase of leisure), there is a feature of actual income taxes which is slighted by such a point of view: it is institutionally feasible to assess taxes at different *rates* on different individuals; in particular, progressive taxation of earnings is possible. This means that the income tax and such related taxes as the expenditure tax are potentially important instruments for meeting distributional objectives. We now turn to the studies which consider the trade-off between such distributional objectives and economic efficiency.

## II. The Optimal Income Tax Literature

The problem of optimal income taxation has a long history in economics.[6] However, most of the recent literature stems from a paper published by James Mirrlees in 1971 A natural way to organize our discussion, then, is to summarize Mirrlees techniques and conclusions, and then view the ensuing literature as an attempt to explain and modify some of his results.

In Mirrlees' model, society is composed

[6] See especially F. Y. Edgeworth's important contribution.

of individuals who have identical atemporal utility functions in after tax income and leisure. Individuals differ only in their earning abilities (wage per hour). The government must collect an exogenously determined amount of tax revenue. The problem is to find an income tax schedule (tax function) which maximizes the sum of individuals' utilities subject to this revenue constraint.

Using the tools of the calculus of variations to solve the constrained maximization problem, Mirrlees finds that the optimal tax function exhibits marginal tax rates between zero and one, and that when it is operative, part of the population does not work. Although these results may seem weak, they are really quite remarkable given the absence of specific functional forms for the key relationships in the problem.

In order to get more specific results, more specific assumptions must be built into the analysis. Mirrlees assumes that the utility functions are Cobb-Douglas, and considers both log normal and Pareto distributions of earnings abilities. With these assumptions, the following results emerge: (a) the optimal tax function is approximately linear with a negative intercept, and (b) the optimal tax function is characterized by "low" marginal tax rates which *fall* somewhat with income. Atkinson's (1973) interpolations of Mirrlees' results indicate rates in the neighborhood of 20 percent.

Mirrlees was surprised at how low the marginal tax rates were: " . . . I must confess that I had expected the rigorous analysis of income-taxation in the utilitarian manner to provide an argument for high tax rates. It has not done so." (A study by Ray C. Fair in the same year also generated fairly low implied marginal tax rates.) Apparently, those who read the Mirrlees paper also found the low marginal tax rates counterintuitive, for much of the literature appears to be an attempt to explain them.

One concern was the maximand of Mirrlees' problem, an unweighted sum of individual utilities, which implies that a "util" to a rich individual adds as much to social welfare as a "util" to a poor individual. To what extent would more egalitarian results (i.e., higher marginal tax rates) emerge if a social welfare function were used which weighted the utilities of the rich less than those of the poor? Atkinson and Feldstein (1973) consider social welfare functions of the form

$$W = \left(\sum U_i^v\right)^{1/v} \qquad v \leq 1. \tag{8}$$

Clearly, when $v=1$, welfare ($W$) is the simple sum of utilities ($U_i$). When $v$ is less than 1, however, it can be shown that a given increment to the utility of a low utility individual adds more to $W$ than if awarded to a high utility individual. It should be noted, however, that the specifications of the social welfare function and the individual utility functions are not really independent of each other. We could, for example, specify the utility of the $i$th individual to be $U_i^v$ and then write social welfare as the arithmetic sum of these utilities.

Atkinson focuses attention on the case in which $v$ approaches minus infinity. Under such circumstances, it can be shown that maximizing $W$ is equivalent to maximizing the utility of the worst off individual in society: the maximin case. This case has received considerable attention due to philosopher John Rawls' argument that it is particularly compelling as an ethical criterion. (A number of criticisms of Rawls' position are suggested by Alvin K. Klevorik.)

Atkinson uses a Rawlsian social welfare function in a model with a linear income tax, no net government revenue requirement (i.e., taxation for redistribution

only), and a Pareto distribution of skills in the economy. He finds that optimal marginal tax rates range between 30 and 45 percent. Thus, one solution to the mystery of Mirrlees' low marginal tax rates is his formulation of the objectives of the government. Social welfare functions which are more egalitarian than the classical utilitarian variety may yield higher marginal rates.

Another potential explanation for Mirrlees' results is the Cobb-Douglas assumption on the form of individuals' utility functions. Stern has investigated this possibility by assuming that individuals have constant elasticity of substitution (*CES*) utility functions in leisure and income. Using results on the elasticity of labor supply from the econometric literature,[7] he finds that an elasticity of substitution of about .4 is more realistic than 1.0.[8] When a variant of Mirrlees' problem is solved using *CES* utility functions with this lower elasticity of substitution, the optimal marginal tax rates are substantially higher—without appeal to a more egalitarian social welfare function.

Reexamination of the social welfare function suggests another possible explanation for the low tax rates typically generated by optimal income tax studies. Our intuition about optimal income taxation may perhaps be conditioned on societal objective functions which are not utilitarian-individualistic. For example, the presence in the social welfare function of a variable parameterizing the "aesthetics" of the income distribution would lead to more egalitarian results.[9] Similarly Feldstein (forthcoming) has shown that if interdependent utility functions are allowed for, very high marginal tax rates may be appropriate.

We now turn to a limitation of the Mirrlees model which is just beginning to receive attention, its atemporal setting. The appropriate taxation of capital income is one of the most controversial aspects of the tax system, yet the studies cited above for the most part ignore it. Janusz A. Ordover and Edmund S. Phelps examine the optimal mix of taxes on two factors of production (capital and labor) in a one sector neoclassical growth model.[10] Their model is very general, and therefore no results on tax rates emerge which can be compared to those discussed above. Moreover, the only social welfare function they consider is the maximin case. Despite these limitations, explicit attention to the taxation of capital income in the optimal income tax framework is an important step which will no doubt stimulate further research.

We could continue to list additional aspects of the Mirrlees model which have been changed and expanded in order to ascertain their effects on optimal tax rates.[11] However, the basic thrust of the literature should now be clear. An exogenously determined amount of tax revenue must be raised by income taxes on individuals whose economic choices are distorted by the presence of those taxes. Given technological and behavioral assumptions, the optimal tax schedule is that which leaves some social welfare function at a maximum after the tax is collected. The literature shows how various

[7] These are measures of the elasticity of hours per year with respect to the wage, and thus do not take into account other, perhaps more important, dimensions of labor supply.

[8] If the elasticity of substitution were zero, lump sum taxation would be possible. If the elasticity of substitution were infinite, no revenue could be raised.

[9] Such a social welfare would be non-Paretian, but there is nothing to prevent a reasonable set of value judgments from allowing for such a possibility.

[10] E. Sheshinski (forthcoming a) considers taxation in a one sector neoclassical growth model with earned and unearned income taxed at the same rate.

[11] For example, Stern has suggested changing the assumptions on the underlying distribution of skills, Sheshinski (forthcoming b) focuses attention on a model in which taxes influence human capital accumulation, and Feldstein (1973) allows for an endogenous wage.

assumptions on these components lead to different conclusions regarding the shape of the optimal income tax schedule.

### III. Concluding Remarks

The accomplishments of the optimal taxation research have been considerable. It has upset many comfortable rules of thumb and lent precision to many informal arguments. But there remains work to be done. Part of this work will, of course, consist of increasing the stock of variations on the basic problems for which solutions have been described. Another, and very important part will consist of the attempt to determine quantitatively which of these problems best describes the actual economy to be taxed—filling in all those empty boxes with real, estimated elasticities.

However, work of another kind is needed to advance the normative power of the analysis. Normatively the optimal tax literature rests on a utilitarian base. It is true that the optimal commodity tax results, or some of them at least, can be cast, in a form which says: if your tax system doesn't look like this, there is a potential bargain which can be struck among your citizens which would make all better off. However, these bargains are complex and their possibility tends to be eliminated by the very assumptions that require the use of second-best instruments in the first place. For practical application implicit interpersonal utility comparisons are required. The optimal income tax results are also dependent on such comparisons. The missing link is a welfare function, and the question is how does one persuade a legislature or an electorate to decide in accordance with some particular welfare function?

Asking the optimal tax researchers to resolve this is effectively to ask them to make welfare economics persuasive, obviously a tall order. But, interestingly, the tax literature has always appealed to non-utilitarian criteria as well.

Thus, missing from the optimal tax literature is the idea of horizontal equity, the notion that " . . . people in equal positions should be treated equally." (Richard A. Musgrave 1959, p. 160) (Customarily, "equal positions" are defined in terms of an observable index of ability to pay such as income, expenditure, or wealth.) In none of the studies discussed above has the injunction to treat equals the same appeared either as a constraint in the maximization problem, or as an argument in the objective function. Therefore, they will in general[12] fail to provide horizontal equity. In light of this, Musgrave (forthcoming) and others have suggested that it is inappropriate to characterize such schemes as "optimal."

Defining horizontal equity in terms of income is inadequate because individuals with identical opportunity sets but different tastes will have different incomes. An alternative way to define equal position would be identical opportunity sets. However, it seems more in the spirit of the optimal taxation literature to define equal position in terms of utilities: individuals are "the same" only if they derive identical amounts of utility from their consumption and leisure bundles. The choice of a criterion for horizontal equity is important because when tastes differ between individuals, different criteria may lead to different conclusions as to the fairness of a given tax. For example, an income tax which is perfectly fair according to conventional notions of horizontal equity hurts an "income lover" more than a "leisure lover."

In an attempt to put the discussion of

[12] It can be shown that if all individuals have identical tastes and there is only one type of ability, then horizontal equity will be satisfied by virtually any broad-based tax. (See Feldstein, forthcoming.) Such assumptions, as we have seen, are built into a number of the optimal tax studies. For an exception, see Diamond and Mirrlees.

horizontal equity and the optimal taxation literature on the same plane, Feldstein (forthcoming) has redefined the principle of horizontal equity in terms of utility rather than ability to pay.[13] However, complete integration of horizontal equity into the optimal tax framework remains to be done. Perhaps this could be accomplished by including some measure of departure from horizontal equity as an argument in the social welfare function, but this approach is bedeviled by conceptual difficulties in measuring departures from horizontal equity.[14]

It may well be that horizontal equity, ancient and honorable criterion of tax policy though it be, is not a helpful concept. However, apparent appeal of this nonoperational idea to practical people suggests the attractiveness of properties of a tax structure which are independent of the economy to which that structure is applied. To discover whether there are any such properties which significantly narrow the range of "good" tax structures might be a useful topic of research.

[13] "If two individuals would be equally well off (have the same utility level) in the absence of taxation, they should be equally well off if there is a tax."

[14] See Rosen for a discussion of these problems and some attempts to surmount them.

## REFERENCES


**A. B. Atkinson,** "How Progressive Should the Income Tax Be?" in Longman's, *Essays on Modern Economics,* London 1973.

**A. B. Atkinson and J. Stiglitz,** "Alternative Approaches to the Redistribution of Income," *J. Publ. Econ.,* forthcoming.

**M. Boiteux,** "Sur la gestion des monopoles publics astreint a l'equilib budgetaire," *Econometrica,* 1956, *24,* 22–40, reprinted in translation by W. J. Baumol and D. F. Bradford, "On the Management of Public Monopolies Subject to Budgetary Constraints," *J. Econ. Theory,* Sept. 1971, *3,* 219–40.

**M. J. Boskin,** "Optimal Tax Treatment of the Family," Stanford Univ., mimeo. 1973.

**W. J. Corlett and D. C. Hague,** "Complementarity and the Excess Burden of Taxation," *Rev. Econ. Stud.,* 1953–54, *21,* 21–30.

**P. A. Diamond and J. A. Mirrlees,** "Optimal Taxation and Public Production: II," *Amer. Econ. Rev.,* June 1971, *41,* 261–78.

**J. Dupuit,** *Traite theorique et pratique de la conduite et de la distribution des eaux,* Paris 1854.

**F. Y. Edgeworth,** "The Pure Theory of Taxation," *Econ. J.,* 1897, *7,* 46–70.

**R. C. Fair,** "The Optimal Distribution of Income," *Quart. J. Econ.,* 1971, *85,* 551–79.

**M. Feldstein,** "On the Optimal Progressivity of the Income Tax," *J. Polit. Econ.,* 1973, *2,* 357–76.

———, "The Theory of Tax Expenditures," Harvard Univ., mimeo. 1975.

———, "On the Theory of Tax Reform," *J. Publ. Econ.,* forthcoming.

**M. Friedman,** "The Welfare Effects of an Income and an Excise Tax," *J. Polit. Econ.,* 1952, *60,* 1–24.

**A. K. Klevorick,** "Discussion," *Amer. Econ. Rev. Proc.,* May 1974, *64,* 158–61.

**I. M. D. Little,** "Direct vs. Indirect Taxes," *Econ. J.,* 1951, *61,* 577–84.

**J. A. Mirrlees,** "An Exploration in the Theory of Optimum Income Taxation," *Rev. Econ. Stud.,* 1971, *38,* 179–208.

**R. A. Musgrave,** *The Theory of Public Finance,* New York 1959.

———, "Optimal Taxation, Equitable Taxation and Second-Best Taxation," *J. Publ. Econ.,* forthcoming.

**J. A. Ordover and E. S. Phelps,** "Linear Taxation of Wealth and Wages for Intragenerational Justice: Some Steady-State Cases," *Amer. Econ. Rev.,* Sept. 1975, *65,* 660–73.

**F. P. Ramsey,** "A Contribution to the Theory of Taxation," *Econ. J.,* 1927, *37,* 47–61.

**J. Rawls,** "Some Reasons for the Maximin Criterion," *Amer. Econ. Rev. Proc.,* May 1974, *64,* 141–46.

**H. Rosen,** "Income, Utility, and Horizontal Equity Under the U.S. Income Tax," Princeton Univ., mimeo. 1975.

**P. A. Samuelson,** "Theory of Optimal Taxation," unpublished, approx. 1952.

A. Sandmo, "A Note on the Structure of Optimal Taxation," *Amer. Econ. Rev.*, Sept. 1974, *64*, 701–06.

E. Sheshinski, "Income Taxation and Capital Accumulation," *Quart. J. Econ.*, forthcoming, a.

———, "On the Theory of Optimal Income Taxation," *J. Publ. Econ.*, forthcoming, b.

J. S. Stamp, *Fundamental Principles of Taxation in the Light of Modern Developments*, London 1921.

N. Stern, "On the Specification of Models of Optimum Income Taxation," *J. Publ. Econ.*, forthcoming.

# On Some Recent Econometric Research in Public Finance

By Michael J. Boskin*

In a field like public finance which is intimately concerned with the formulation and evaluation of public policy, the continuing search for scientific information (upon which to base such policies) acquires a particularly large marginal social product. Yet, throughout the history of public finance there has been at times a harmonious blend of the analytical with the empirical and at other times a flagrant clash between them. From such clashes a synthesis frequently has been developed which gains widespread acceptance. The point is simply that scientific information is only acquired as the intersection of theoretical and empirical research. Once this is recognized, there are three basic ways in which we are able to improve our basic knowledge: we may develop new and better theoretical insights; we may obtain more data, or improved data, or both; and we may develop superior methods by which to analyze data.

In the last few years, a number of excellent examples of each type of progress in public finance have been produced. The purpose of the present paper is to discuss several examples of recent econometric work in order to indicate the enormous progress that has been made in obtaining vastly improved estimates of some of the key parameters with which we are concerned. In so doing, we shall also note how these empirical findings sometimes force us to abandon previously held beliefs and ways of thinking about certain problems.

Toward this end, Section I discusses some recent evidence on the effect of the charitable contributions deduction on giving to charity. This recent work is based on a variety of data sources and employs some relatively advanced econometric techniques to estimate the price and income elasticities of charitable giving. The estimated price elasticities are much larger than those which had been accepted previously and have strikingly different implications for public policy. Thus a reworking of what is basically a simple problem of estimating the demand for a single commodity has resulted in an enormous advance. Section II deals with recent advances in the analysis of the effects of taxes on the supply of labor. I include this topic both to illustrate my major point in an area of incredibly rapid technical progress and to demonstrate how public financiers also must rely on research from related fields. The conceptual and estimation problems in this area are much more complex than those for the analysis of tax incentives for charitable giving, but we are converging (slowly) on a better understanding of the ways in which taxes affect labor supply.

Section III reports on some recent research on the effects of taxes on saving behavior. Again, more and better data and advances in econometric techniques have led to empirical results which are forcing

* Stanford University and the National Bureau of Economic Research. This paper is part of a planned larger survey of empirical research in public finance. It is meant to illustrate the types of progress being made, not to survey the field comprehensively. Indeed, anything approaching a complete list of references would probably exceed my modest page limit! See G. F. Break (1974) for a partial survey.

us to rethink some widely accepted propositions on the incidence and effects of taxation. Here the focus is on simple macroeconomic models and different types of data and estimation problems than in the examples cited above.

Finally, Section IV offers a brief summary and suggestions for future research.

## I. The Effects of Tax Incentives on Giving to Charity

As an example of a wide range of problems in the analysis of the effects of tax deductibility of any particular item, consider the deductibility of charitable contributions in the income and estate taxes. For those persons itemizing deductions, deductibility lowers the cost of giving to $1 - r$, where $r$ is the marginal tax rate. A frequently used measure of the efficiency of a tax deduction is the ratio of the induced increase in a particular activity (such as charitable giving) due to the deduction, to the revenue foregone by the Treasury. Obviously, this ratio is a key parameter in policy discussion and its size is a function of the price elasticity of demand for the commodity under consideration.

Until quite recently, it was thought that the charitable deduction had very little impact on giving. For example, William Vickrey (1962) concludes that ". . . the evidence would seem to indicate that while the deductibility may increase the gross amount of contributions, it does so by less than the tax relief granted." Further, Michael Taussig (1967) estimated that charities received only five cents for every dollar of revenue foregone by the Treasury. Unfortunately, Vickrey does not estimate a demand for giving equation and Taussig does so in a way that obscures the relationship between giving and after-tax cost.[1] Thus, their conclusions are not the sort of information upon which one would want to base tax policy toward charitable contributions.

In somewhat of a tour de force, Feldstein (1975a,b) and Feldstein and his collaborators (1975a,b, 1976) have examined several complementary data sources and estimated demand for charitable giving equations. The basic equation estimated in each of these studies is of the form

$$ln\ G_i = \alpha_0 + \alpha_1\ ln\ Y_i + \alpha_2\ ln\ P_i + \epsilon_i$$

where $G$ is charitable contributions, $Y$ is income and $P$ is net-of-tax price. The results indicated a price elasticity in excess of unity in virtually every case. Note that this result suggests that charitable organizations gain absolutely *more* than the Treasury loses in revenue, i.e., the deduction is more than fully efficient. Feldstein (1975a) also notes that eliminating the deduction would increase inequality in the after tax and giving distribution of income.

The most important thing to note about this work from a methodological point of view is the bringing to bear of a variety of different data sources to attempt to answer the same question: How large is the price elasticity of giving to charity? Feldstein and his collaborators use pooled aggregate statistics of income data by income class, individual tax return data from 1962 and 1970 tax returns and household survey data. Thus the advantages of several types of data have been exploited. For example, the household survey data include much information on demographic variables which influence charitable giving and these have been used as additional regressors in the basic equation; the pooled time series data are useful in dealing with the permanent-transitory income problem, etc.

Contrast this approach of examining several data sources in analyzing a particular problem with the typical approach in

[1] See Martin Feldstein, 1975a.

empirical work in public finance (and other fields of economics). Usually, we witness a sporadic series of difficult to compare studies produced over a long period of time using different data sources, econometric techniques and definitions of variables. Then someone writes a survey paper bemoaning the differences in the studies and heroically attempting to determine if the difference is due to the different data sources, the different techniques used to analyze the data or some change over time or across demographic groups in the underlying behavior. The policy maker frequently is left to choose from a bewildering array of conflicting studies.

While applying the same techniques to a variety of types of data sources will not always lead to virtually identical results, it frequently will be the case that knowledge of the advantages and limitations of the alternative data sources can be combined with econometric techniques to narrow substantially the confidence intervals we attach to estimates of key parameters.

A second major methodological feature of these studies is the attempt to examine a variety of functional forms, definitions of the variables and subsamples of the population in order to test the robustness of the basic results. For example, in addition to the basic constant elasticity demand equation described above, Feldstein et al. estimated equations in which the price elasticity was allowed to vary with price and the income elasticity with income; in which separate price elasticities were estimated for separate income classes and age and marital status groups; and in which alternative reasonable measures of price and income were used. Fortunately, the results supported the basic conclusion of a greater than unitary price elasticity.

Economic theory *generally* guides us to a set of structural relationships to study; it *sometimes* places meaningful restrictions on some of these relationships (nonnegativity, symmetry, etc.); however, it *rarely* implies a precise functional form. Further, the available data usually do not contain information which enables us to conform exactly to our theoretical definitions of variables. Hence, it is important to explore the robustness of our results to alternative specifications. Combined with an analysis of alternative data sources, this approach in at least a subset of cases will lead to results which may be imbedded in the subsequent analysis and discussion of the problem under investigation. For example, the results of Feldstein et al. have replaced the earlier results and the simple cross-tabulations from tax data which have ignored these behavioral responses.

We shall see below that studies of the effects of taxes on labor supply suffer in the extreme from noncomparability of data source, estimation technique, functional form and variable measurement, whereas studies of saving behavior partly follow the pattern outlined above. Before turning to these studies, however, it is worth noting one more type of advance that has been made in the recent literature on taxes and charitable contributions.

A variety of recently developed econometric techniques are continually in the process of diffusion throughout the profession. One type which is becoming increasingly important in public finance and other fields which analyze data on individual households is the analysis of qualitative and limited dependent variables. These techniques are designed to deal with situations where the dependent variable in a relationship must take on one of a mutually exclusive set of discrete values or piles up at some limit. For example, a person chooses whether or not to be in the labor force, which college to attend, which transport mode to take to work, etc. In the an-

alysis of data on charitable bequests, one is confronted with precisely such a problem. Only twenty percent of decedents leave anything to charity. Hence, in regressions such as those described above for charitable contributions under the income tax, of charitable bequests on wealth and price, eighty percent of the observations on the dependent variable are zeroes! Fortunately, techniques have been —and are being—developed to handle these problems. As an example of the potential difficulties involved, Stephen K. McNees (1973) estimated such an equation using ordinary least squares. His results predict that most estates bequeath negative amounts to charity.[2] I have reestimated his equations using the truncated normal regression model and have estimated substantial price elasticities of charitable bequests. We shall return to a discussion of these techniques below.

## II. The Effects of Taxes on the Supply of Labor

The importance of estimates of the effects of taxes on the supply of labor to the making of intelligent tax policy is difficult to overemphasize. Answers to such structural questions as an appropriate unit of account for taxation and to such fundamental issues such as the optimal progressivity of the income tax depend crucially upon the effects of taxes on labor supply. Indeed, much recent theoretical and empirical research in public finance focuses on relaxing the assumption of fixed factor supplies. We deal with labor and capital in turn.

The last few years have witnessed an explosion of research on labor supply. It illustrates two other crucial types of (admittedly sporadic) progress being made in econometric research in public finance: the availability of new bodies of data potentially useful in the study of a wide variety of problems and major advances in econometric technique. New and important sets of data containing information on individual households have become increasingly available. This development in itself has enormous import for empirical work in public finance as it potentially frees us from exclusive reliance on time series data and its inherent problems. Examples include the 1966 and 1967 *Surveys of Economic Opportunity*, the *Panel Study of Income Dynamics*, and the *National Longitudinal Surveys*. These data allow us to follow the same households over several years; special methods for analyzing such data are now emerging. Finally, controlled social experiments are producing data which complement these other data sources in important ways.

All of these data sources have been used in attempts to estimate labor supply functions. Unfortunately, different techniques have been applied in each of these studies. For example, Boskin (1974) and Robert E. Hall (1974) use the Survey of Economic Opportunity to estimate labor supply for various groups as a function of after-tax income and wage rates. James J. Heckman (1975), using the National Longitudinal Survey, makes an important conceptual advance, but totally ignores taxes in measuring wage rates! Harvey S. Rosen (1973) reports results from this same body of data supporting the hypothesis that it is after-tax rather than gross wage rates that matter. It seems as if each paper has made an advance but generally ignores those made previously. Hence, the studies are extremely difficult to compare and no clear consensus on elasticities of labor supply has emerged. About all that can be said in surveying the studies is that the labor supply of wives seems to be more wage-elastic than that of husbands; while the latter (at least ignoring human investment on the job) appears to

[2] See Boskin, 1976.

be relatively inelastic, estimates of the former range from well under one to well over two. Worse yet, every estimate I have seen is inconsistent in the statistical sense due either to econometric technique used or to measurement of the variables.

The study of labor supply also provides an excellent example of the interaction of data availability, policy relevance and conceptual advances. There is no question that the intellectual interest (not to mention financial support) in research on labor supply was stimulated by the debate over the labor supply effects of a negative income tax. A variety of successive and important conceptual advances have been made. For example, the typical labor supply study of several years ago merely regressed hours of work on wage rates and a measure of income. For groups of the population with labor force participation rates well below 100%, this procedure is inappropriate. Further, imputing wages to those not currently working by comparison with those who are may also lead to inconsistent estimates.[3] For example, Heckman points out that if

$$W_M = f(E, S)$$

$$W_A = f(C, A, W_s, H)$$

(where $W_M$ is the market wage rate facing a potential worker, $W_A$ is the asking wage, $E$ is labor market experience, $S$ is years of schooling, $C$ is the number of children, $A$ is assets, $H$ is hours of work, and $W_S$ is spouse's wage), hours of work will adjust so that $W_A = W_M$ for workers, whereas $W_M < W_A$ if hours of work in the market are zero. He uses this simple observation to build an econometric technique for estimating these two functions, which may be solved for the market labor supply function. The technique is a generalization of the truncated regression model (usually called Tobit analysis) discussed above. It solves many of the problems encountered in the previous work on labor supply.

Unfortunately, Heckman's estimates are not yet readily usable because he ignores taxes in the measurement of wage rates; hence his estimates are inconsistent. G. Hanoch (1975), in an excellent paper, has extended the Heckman notion in a variety of ways. However, nothing like a systematic combination of the improvements in measurement, in conceptualization, and in econometric technique has been used to analyze a variety of complementary data sources. From the point of view of the consumers of estimates of the effects of taxes in labor supply, such a synthesis is badly needed and long overdue.

### III. Taxation and Saving

The recent econometric research in taxation and saving behavior has not been pushed forward primarily by new econometric techniques. Rather, it has been, marked by a more careful examination of the fundamental relationships, an attempt to incorporate potentially important missing elements and an examination of alternative data sources and functional forms. In addition to the obvious direct usefulness of this work, it is particularly important because the recent results are forcing us to reexamine widely held views on tax incidence.

The three most important lines of inquiry, in my opinion, have been the study of the relationship of corporate and personal saving, the social security system, and the interest elasticity of the saving rate and public and private marginal propensities to save.

The typical consumption or saving function related consumption or saving to private income after taxes, excluding retained earnings. Feldstein (1973) and Feldstein and G. Fane (1972) have extended this

[3] See J. Heckman, 1975.

model to include the "components" of capital gains, including retained earnings. The estimates based on *U.S.* and British time series data, respectively, suggest that the marginal propensity to consume out of retained earnings is two-thirds that out of disposable income. Again, these results are derived from a variety of specifications of the basic relationship, including alternative measures of permanent income. They suggest that consumers see through the corporate veil and (partially) adjust their personal saving to changes in corporate saving.

A second line of work, also pioneered by Feldstein (1974a, 1975d), and contributed to by Alicia H. Munnell (1974), has been the study of the effects of social security on private saving. The enormous increase in social security benefits financed on a pay-as-you go basis has been one of the most important features of the postwar *U.S.* economy. It is conceivable that the expectation of social security payments during retirement, financed by the generation then working, leads to a decrease in private saving. Feldstein has examined two different types of data to attempt to answer this question. His studies of aggregate *U.S.* time series and an international cross section imply that an enormous substitution of expected social security benefits for private saving has occurred. Indeed, his best estimates are that the substitution is virtually dollar for dollar and that the *U.S.* private saving rate has been reduced by thirty percent or more by the social security program.

Finally, Boskin (1976) has reexamined the questions of the interest elasticity of the saving rate and differential marginal propensities to save publically and privately raised by Feldstein (1974a,b), Peter A. Diamond (1970), C. Wright (1969) and Warren E. Weber (1970, 1975). Paying particularly close attention to the measurement of saving and real after-tax rates of return, he estimates an interest elasticity of approximately .4; this estimate is much larger than that found in previous work. He also concludes that the marginal propensity to save out of private after-tax income is over twice that out of government revenue.

Again, as with the study of charitable contributions, these studies are marked by an attempt to bring several data sources to bear on the same question; to examine the robustness of the conclusions to alternative sample periods, definitions of the variables, and functional forms; and to apply the most reasonable econometric techniques warranted by the problem and the data. Unlike the studies of charitable contributions, the studies of taxation and saving have implications far beyond the particular problem addressed. For example, some widely held notions on the incidence of taxation *assume* that the policies under study do not affect saving, the capital-labor ratio, and hence the functional distribution of income. Indeed, much of the profession seems to give a behavioral interpretation to what has come to be called Denison's Law.[4] The studies mentioned above call this assumption into serious question. Feldstein's estimates of the impact of the social security system on private saving indicates that the *benefits* side of social security has sharply increased the rate of interest and decreased the wage rate. My estimates of the interest elasticity of the saving rate imply that the traditional view—embodied in virtually all empirical estimates of tax incidence—that a proportional income tax is borne in proportion to income is erroneous; labor bears more than its proportion of national income. Further, taxes on capital income are largely shifted onto labor; hence, the traditional views on the inci-

[4] See P. David and J. Scadding, 1974.

dence of capital income taxes may need to be replaced.

## IV. Conclusion

I have tried to illustrate in the brief space provided me the important advances that have occurred recently in econometric work in public finance. It seems to me we have entered a new era where both the quality and quantity of econometric research are improving rapidly. As noted above, there are several reasons for this state of affairs. First, a variety of new data sources have become available. Second, new techniques for analyzing that data have been developed. Third, a style of research placing heavy emphasis on exploiting alternative data sources and examining the robustness of the results to alternative specifications has emerged.

I have no doubt that while the particular studies mentioned will be improved by still more and better data and techniques, the methodological message they impart will serve to upgrade the quality of both the future empirical research in public finance and the information upon which sound public policies ultimately must be based.

## REFERENCES

**M. Boskin**, "The Economics of the Labor Supply," in G. Cain and H. Watts, eds., *Income Maintenance and Labor Supply,* New York 1974.

———, "Estate Taxation and Charitable Bequests," *J. Publ. Econ.,* Feb. 1976.

———, "Taxation Saving and the Rate of Interest," National Bureau of Economic Research, Jan. 1976.

**G. F. Break**, "Taxation," in A. Blinder, et al., *The Economics of Public Finance,* Brookings Inst., Washington 1974.

**P. David and J. Scadding**, "Private Savings: Ultrarationality, Aggregation and Denison's Law," *J. Polit. Econ.,* Apr. 1974.

**P. Diamond**, "Incidence of an Interest Income Tax," *J. Econ. Theory,* Sept. 1970.

**M. Feldstein**, "Tax Incidence with Growth and Variable Factor Supply," *Quart. J. Econ.,* Nov. 1974a.

———, "Incidence of a Capital Income Tax with Variable Savings Rates," *Rev. Econ. Stud.,* Aug. 1974b.

———, "The Income Tax and Charitable Contributions: Part I," *Nat. Tax J.,* Mar. 1975a.

———, "The Income Tax and Charitable Contributions: Part II," *Nat. Tax J.,* June 1975b.

———, "Social Security, Induced Retirement and Aggregate Capital Accumulation," *J. Polit. Econ.,* Oct. 1974c.

———, "Social Security Programs and Private Savings: International Evidence in an Extended Life-Cycle Model," Harvard Institute for Economic Research, disc. pap., 1975c.

———, "Tax Incentives, Corporate Savings and Capital Accumulation in the United States," *J. Publ. Econ.,* Apr. 1973.

——— **and G. Fane**, "Taxes, Corporate Dividend Policy and Personal Savings: the British Postwar Experience," *Rev. Econ. Statist.,* 1974.

——— **and C. Clotfelter**, "Tax Incentives and Charitable Contributions in the United States: A Microeconometric Analysis," *J. Publ. Econ.,* Feb. 1976.

——— **and A. Taylor**, "Taxation and Charitable Contributions: An Analysis of Individual Tax Return Data for 1962 and 1970," Harvard Institute for Economic Research, disc. pap., 1975a.

——— **and M. Boskin**, "Effects of the Charitable Deduction on Contributions by Low and Middle Income Households: Evidence from the National Survey of Philanthropy," CREG memo no. 192, Stanford Univ., 1975b.

**R. Hall**, "Wages, Income and Hours of Work," in G. Cain and H. Watts, eds., *Income Maintenance and Labor Supply,* New York 1974.

**G. Hanoch**, "Hours and Weeks in the Theory of Labor Supply," Rand memo, 1975.

**J. Heckman**, "Shadow Prices, Market Wages, and Female Labor Supply," *Econometrica,* 1974.

S. McNees, "Deductibility of Charitable Bequests," *Nat. Tax J.*, Mar. 1973.

A. Munnell, "The Impact of Social Security on Personal Savings," *Nat. Tax J.*, Dec. 1974.

H. Rosen, "The Impact of U.S. Tax Laws on the Labor Supply of Married Women," unpublished dissertation, Harvard 1973.

M. Taussig, "Economic Aspects of the Personal Income Tax Treatment of Charitable Contributions," *Nat. Tax J.*, Mar. 1967.

W. Vickrey, "One Economist's View of Philanthropy," in F. Dickinson, ed., *Philanthropy and Public Policy,* National Bureau of Economic Research 1962.

W. Weber, "The Effect of Interest Rates on Aggregate Consumption," *Amer. Econ. Rev.*, Sept. 1970.

———, "Interest Rates, Inflation, and Consumer Expenditure," *Amer. Econ. Rev.*, Dec. 1975.

C. Wright, "Saving and the Rate of Interest," in A. Harberger and M. Bailey, *The Taxation of Income from Capital,* Brookings Inst., Washington 1967.

STATE AND LOCAL PUBLIC FINANCE

# Institutional Constraints and Local Community Formation

*By* RICHARD E. WAGNER*

The interest of economists in alternative patterns of urban government stems primarily from the publication of papers by Charles M. Tiebout and George J. Stigler. While these papers were essentially positive in orientation, subsequent work explored the normative properties of various models of urban government. These normative models, which have been surveyed exhaustively by Kenan Bulutoglu, have predominantly focused on the external costs that units of government impose on one another. While particular models vary, the general framework is quite uniform. A model is postulated in which the activity of one jurisdiction influences positively the utility level of residents of other jurisdictions. The analyst then postulates an "independent adjustment equilibrium," and concludes that this equilibrium is inefficient because individual jurisdictions fail to take into account the impact of their actions on other jurisdictions.

Albert Breton, Wallace E. Oates, Mancur Olson, and Gordon Tullock attributed this nonoptimality to the inappropriate specification of boundaries, for the benefit range of the publicly provided service fails to coincide with the boundary of the jurisdiction providing the service. Once it is suggested that some boundaries are better than others, the principles of community formation or boundary selection come to the fore of the economic analysis of local government. Optimality models of club or community formation have been developed by such authors as Martin C. McGuire, A. Mitchell Polinsky, and Hirofumi Shibata, these various works all being inspired by James M. Buchanan's (1965) theory of clubs.

This paper advances the thesis that the economic analysis of local government and community formation should deemphasize the construction of normative models, and should concentrate instead on the positive analysis of alternative institutional arrangements. The first section of this paper argues that the economic analysis of urban government should become less "neoclassical" and more "institutionalist" because optimality models can never be related to social reality *independently of the institutional framework within which collective choices emerge*. The second section moves beyond methodological exhortation to the description of some institutional research, with this description serving to reinforce the theme of the first section.

## I. Mind, Matter, and Institutionalist Themes

The prescriptive use of optimality models of local government requires a framework in which the analyst knows *by assumption* the mental states of the participants in the economic process. Once the essential privacy of individual minds is recognized, however, the analyst can no longer adhere to such a framework. As Friedrich A. Hayek (1937, 1945) has

* Professor of Economics, Virginia Polytechnic Institute and State University.

argued with especial clarity and force, the inherent subjectivity of the data of the models prevents an external observer from relating his conceptual constructions to reality independently of actual choices, with particular choices themselves dependent upon the constraints created by particular institutional settings.

Consider, for instance, McGuire's suggestion that institutional rules facilitating easy community formation might not necessarily be optimal because an outcome analogous to a Chamberlinian tangency solution might result. Consequently, too many communities would exist in the sense that local public services would be provided more expensively than necessary. Following this suggestion, suppose an analyst were to judge that public services would be provided most cheaply with jurisdictions of say, 50,000 members, but observes that actual jurisdictions contain, say, 30,000 members. What inference should the analyst draw from this discrepancy between optimal size and actual size? Should the situation be called "nonoptimal," or should the presumption about optimal size be considered refuted?

The analytical proposition that the least-cost size is 50,000 members is, of course, nothing but an economist's representation of the presumed mental states of the participants in the economic process. This representation, if accurate, implies that a resident of a jurisdiction with 30,000 members would sense a personal disequilibrium. To say that an individual bears a higher cost when jurisdictions contain 30,000 members than when they contain 50,000 members is to say that this individual places a higher value on those options he must sacrifice to maintain a 30,000 jurisdiction than he places on those options he would sacrifice by switching to a jurisdiction with 50,000 members.

But if this is an accurate representation of mental states, how can jurisdictions with 30,000 members survive? Would not the survival of such jurisdictions refute the presumption of inefficiency and suggest instead that the economist had misrepresented the mental states of the participants? If the economist had developed an accurate representation, the choices actually made by the participants would necessarily have corresponded to the person's sensed least-cost option.[1]

Benefit spillovers raise the same set of issues. Suppose, for instance, that a set of governments in an urban area fail to engage in contractual relations of any form. Does the absence of such relations indicate nonoptimality as described by the standard formulation of independent adjustment equilibrium, or does it indicate only that the economist's representation of the mental states of the participants is incorrect?

If the economist's representation were correct, the so-called "independent adjustment equilibrium" characterizes a disequilibrium setting. In such a setting, an incentive exists for units of government to engage in mutually profitable exchanges, and such incentives will persist so long as the situation described by the divergence between the independent adjustment equilibrium and the optimal outcome continues to describe accurately the mental states of the participants. One possible interpretation, then, of the observed absence of contractual relations is that the postulated spillover effects are nonexistent.

Economics, as G. L. S. Shackle has noted, deals with thoughts, not with things.[2] Merely to observe a physical pres-

[1] The necessity of relating concepts of cost to the choices actually made by participants in the economic process and the impossibility of measuring cost independently of choice are developed in James M. Buchanan (1969).

[2] This is aptly expressed by the sentiment: "Economics is a branch of the study of man, and is *essentially* concerned with the nature and processes of his mind" (p. 92, Shackle's italics).

ence, then, cannot be taken as verifying the existence of spillovers. It is not physical states, but mental states with which we deal, and the latter cannot be observed directly but can at best only be inferred through observation of choices within particular institutional settings. The analyst once again merely concocts a representation of the mental states of the participants. But information as to the aptness of any particular concoction can be gained only by observing choices that actually emerge, and these choices will be constrained by the particular features of the institutional environment. The development of rationales for state activity, then, which has been the objective of the normative theorizing, is a quite different activity than the construction of an interpretation of social reality.

Compulsory public education, to select a different illustration, has often been rationalized on the basis of externalities, it being alleged that the educational level of any one person affects positively the utility level of all other persons. E. G. West, however, has developed a coherent, alternative interpretation of the emergence of compulsory schooling as an instrument for enhancing the monopoly position of the educational establishment. Relatedly, Robert J. Staaf has described how in the absence of externalities public education can nonetheless emerge through the operation of majority rule budgetary processes under certain distributions of income and tax shares. The development of a rationale, then, begs the question of the selection of an interpretation of reality.

## II. Rationalization, Refutation, and Community Formation

The customary normative analysis of local government is necessarily incapable of explaining what is at stake in choosing among alternative institutional arrangements for guiding the formation of communities and the selection of boundaries—and it is the institutional arrangements, not the "optimal" allocations themselves, that are subject to choice in the first place. This section reinforces the methodological point developed in the preceding section, and does so in part by illustrating once again the lacuna inherent in the normative approach to urban government.

There may indeed be reasons why mutually profitable trades might fail to be consummated, though the acknowledgment that this may be so shifts the analytical focus from the manipulation of normative models to the examination of alternative institutional arrangements. I, for instance, have described how the incentive to seek out and consummate mutually profitable agreements among governments will vary with such institutional features as the attenuation of ownership rights and the range of services provided by a single unit of government. In a related vein, the Governor's Task-Force on Local Government Reform (California) has noted how restrictions or prohibitions on the ability of people to secede from existing jurisdictions, and to form new ones, would act as barriers to entry, thereby altering the inferences that one might draw from observations regarding the survivability of specific practices or outcomes.

Enormous variety exists among states in the constraints placed upon the formation of local communities, and these constraints pertain to such matters as incorporation and disincorporation, annexation and disannexation, and the creation of special districts. With regard to annexation, for instance, many states require separate majority votes in the annexing city and in the territory to be annexed, while in some states a city unilaterally can annex unincorporated territory. Similar variations exist with respect to municipal incorporation. In eleven states (Connecticut, Delaware, Florida,

Georgia, Hawaii, Maine, Massachusetts, New Hampshire, New Jersey, Rhode Island, and Vermont), municipal corporations are formed by special acts of the state legislature. In the remaining states, general rules of incorporation apply either to all cities or to classes of cities. Some states even give cities the right to amend their own charters.

One of the features of particular interest is the restrictiveness of the constraints placed upon the formation of new communities. While many states permit what may be called free formation of local communities from unincorporated territory, seven states (Alaska, California, Kansas, Maryland, Minnesota, Tennessee, Virginia, and Wisconsin) tightly constrain entry by essentially giving existing localities control over the formation of new communities.

A normative model could, of course, be constructed to rationalize either practice. One might postulate, for instance, that the creation of a new jurisdiction imposes costs on other jurisdictions, and conclude that optimality requires that existing communities be given a voice in the creation of new communities. Sho Sato and Arvo Van Alstyne, for instance, oppose provisions for free entry on the grounds that the use of such provisions "tends to frustrate any rational boundary adjustment" (p. 45). One could also rationalize free entry, perhaps by arguing that it acts as potential competition to promote efficiency among existing units of government. Moreover, the agile normative theorist could even rationalize both rules simultaneously: he would simply suggest that free entry be used when the marginal gains from the promotion of competition exceed the marginal gains from the internalization of externalities, and that restricted entry be used otherwise.

The construction of rationales, however, begs the question of the appropriate interpretation of social reality. For instance, institutional rules that enable existing communities to restrict the formation of new localities could be interpreted as a practice that diminishes competitive pressures in the public sector. Such rules, then, would operate in much the same manner as would rules that gave existing grocers in some city the right to veto potential new entrants. Since the supply of rationales for some particular activity is limited only by the imagination of the analyst, analytical effort should be focused not on rationalization but on refutation of the presumptions about reality implicit in the various possible rationalizations.

California has operated under both types of institutional regimes. Before 1964 California essentially had free entry, but after 1964 it gave existing communities a veto over the formation of new communities. In each of the 58 counties, Local Area Formation Commissions (LAFCOs) were created; these LAFCOs consisted of representatives from the county and from the incorporated municipalities within the county and were given the right of vetoing the formation of new communities within the county.

While Dolores T. Martin and Wagner estimate that the creation of LAFCOs reduced by 56 percent the rate of formation of new communities, the question of interpretation revolves around the consequences of this curtailment. It could be argued that LAFCOs enabled communities to create jurisdictions of more efficient, lower-cost size. It could also be argued, however, that by reducing the constraining influence of competition in the public sector, LAFCOs allowed the actual unit cost to rise due to the enhanced survivability of X-inefficiency.

Martin and Wagner estimated a model of both total and per capita expenditure for both the pre- and the post-LAFCO periods, with the units of observation be-

ing the 58 California counties. The hypothesis of identical coefficients as between the two periods was rejected at the .05 level for total expenditure, and at the .001 level for per capita expenditure. Expenditure behavior, then, would seem to differ significantly as between the pre- and the post-LAFCO periods. Furthermore, application of the pre-LAFCO structural relationship to the post-LAFCO data to generate a prediction of what would have happened had LAFCO not been instituted suggests that LAFCO was responsible for a 17 percent increase in total expenditure and a 13 percent increase in per capita expenditure. Theodore Bergstrom and Robert D. Goodman and Thomas E. Borcherding and Robert T. Deacon find that the demand for municipal services is price inelastic, being in the vicinity of –.5. The expenditure consequences of LAFCO in conjunction with the inelastic demand for public services is consistent with the monopoly enhancing interpretation of LAFCO.

I recognize, of course, that particular results are only suggestive, not conclusive, and that there is much room for the comparative examination of alternative institutional arrangements.[3] In juxtaposition with the growing body of positive, interpretative literature on the monopolistic tendencies of governmental institutions, however, this particular result both reinforces and is reinforced by this literature.[4] But transcending issues of particular interpretation, to repeat, is the desirability of a shift in analytical emphasis from the continued spinning of normative, optimality tales to the construction of interpretations of social reality.[5]

[3] Such an institutionalist thrust informs the array of studies being undertaken by Elinor and Vincent Ostrom and others through the "Workshop in Political Theory and Policy Analysis" at the University of Indiana.

[4] For recent empirical studies of this theme, see William A. Niskanen and Wagner and Warren E. Weber.

## III. Concluding Summary

The economic analysis of local government has so far developed primarily as an institution-free collection of abstract models attempting to explore and expound various possibilities. Possibility, however, should not be confused with probability or reality, and I would suggest that the continued trituration of abstract, institutionless optimality models should be replaced by the positive analysis of alternative institutional frameworks. We, as a society, cannot choose allocations directly but can only choose institutions, and it is through these institutions that allocations emerge. It is the institutions and not the allocations, then, that should be our basic elements or building blocks, the primitives in our analytical system. In conclusion and summarization, I would second J. de V. Graaff's conclusion about normative or optimality theorizing: "I do feel very strongly that the greatest contribution economics is likely to make to human welfare, broadly conceived, is through *positive* studies—through contributing to our understanding of how the economic system actually works in practice—rather than through normative welfare theory itself" (p. 170, Graaff's italics).

[5] I am presently preparing a book-length treatment of the organization of government, written from the perspective adopted in this paper, and tentatively titled *Property Rights, Public Choice and Urban Government*.

## REFERENCES

Advisory Commission on Intergovernmental Relations, *The Problem of Special Districts in American Government*, Washington 1964.

T. E. Bergstrom and R. P. Goodman, "Private Demands for Public Goods," *Amer. Econ. Rev.*, June 1973, *63*, 280–96.

T. E. Borcherding and R. T. Deacon, "The Demand for the Services of Non-Federal Governments," *Amer. Econ. Rev.*, Dec. 1972, *62*, 891–901.

A. Breton, "A Theory of Government Grants," *Can. J. Econ. and Pol. Sci.*, May 1965, *31*, 175–87.

J. M. Buchanan, *Cost and Choice*, Chicago 1969.

———, "An Economic Theory of Clubs," *Economica*, Feb. 1965, *32*, 1–14.

K. Bulutoglu, "Fiscal Decentralization: A Survey of Normative and Positive Contributions," manuscript, Istanbul 1974.

Committee for Economic Development, *Reshaping Government in Metropolitan Areas*, New York 1970.

Governor's Task Force on Local Government Reform, *Public Benefits from Public Choice*, Sacramento 1974.

J. de V. Graaff, *Theoretical Welfare Economics*, Cambridge, Mass. 1957.

F. A. Hayek, "Economics and Knowledge," *Economica*, Feb. 1937, *4*, 33–54.

———, "The Use of Knowledge in Society," *Amer. Econ. Rev.*, Sept. 1945, *35*, 519–30.

M. McGuire, "Group Segregation and Optimal Jurisdictions," *J. Polit. Econ.*, Feb. 1974, *82*, 112–32.

D. T. Martin and R. E. Wagner, "The Institutional Framework for Community Formation: An Economic Analysis of Local Agency Formation Commissions in California," manuscript, Blacksburg 1976.

W. A. Niskanen, "Bureaucracy and the Interests of Bureaucrats," *J. Law Econ.*, forthcoming.

M. Olson, Jr., "The Principle of 'Fiscal Equivalence': The Division of Responsibilities Among Different Levels of Government," *Amer. Econ. Rev. Proc.*, May 1969, *59*, 479–87.

W. E. Oates, *Fiscal Federalism*, New York 1972.

A. Polinsky, "Collective Consumption Goods and Local Public Finance Theory: A Suggested Analytic Framework," in *Issues in Urban Public Finance*, Saarbrucken 1973, 166–81.

S. Sato and A. Van Alstyne, *State and Local Government Law*, Boston 1970.

G. L. S. Shackle, *Time in Economics*, Amsterdam 1958.

H. Shibata, "Group Consumption and Marginal Cost Pricing," manuscript, Lexington 1975.

J. Staaf, "Public Education: A Question of Externalities or Income Redistribution," manuscript, Blacksburg 1975.

G. J. Stigler, "The Tenable Range of Functions of Local Government," Joint Economic Committee, *Federal Expenditure Policies for Economic Growth and Stability*, Washington 1957, 213–19.

D. M. Tiebout, "A Pure Theory of Local Expenditure," *J. Polit. Econ.*, Oct. 1956, *64*, 416–24.

G. Tullock, "Federalism: Problems of Scale," *Public Choice*, Spring 1969, *6*, 19–29.

R. E. Wagner, *The Fiscal Organization of American Federalism*, Chicago 1971.

——— and W. E. Weber, "Competition Monopoly, and the Organization of Government in Metropolitan Areas," *J. Law Econ.*, forthcoming.

E. G. West, "The Political Economy of American Public School Legislation," *J. Law Econ.*, Oct. 1967, *10*, 101–28.

# On the Theory of Clubs

*By* EITAN BERGLAS*

In a well-known paper J. M. Buchanan develops a general theory of clubs. He suggests that whenever the utility derived by an individual from a specific good or service is dependent on the size of the consumption group, then a club organization will supply the service efficiently while the market will not.

In this paper (Section I) we show that given the above characterization of commodities (which we regard as an important addition to the economic literature) no market failure need occur. More specifically, taking Buchanan's example, competitive firms can provide swimming pool services as efficiently as clubs.[1] Thus, Buchanan's paper does not provide a sufficient explanation for the creation of clubs.

In Section II, we extend the Buchanan model by allowing the individual to vary the number of times per season he visits the swimming pool. Once this is taken into consideration, it can be shown that providing the service on the basis of either a membership fee or a two-part tariff is inefficient; the optimal pricing system is to charge a price per visit. This will be the price when the service is provided by competitive firms.

Introducing heterogenous tastes (Section III), it is shown that individuals with different tastes will prefer to join separate clubs—and that when the service is provided by the market, the market will provide (and the consumers will prefer) different qualities of service for different types of individuals. In order to allow for the case where individuals with different tastes use facilities providing the same quality of service, it is necessary to bring transportation costs into the model—thus introducing the very complex considerations of spatial equilibrium. The problems associated with this case are discussed and partial solutions are offered. In Section IV we offer some concluding remarks, clarifying the connection with the literature on externalities and increasing returns to scale.

## I

We start by reconstructing Buchanan's model, simplifying its presentation in a way which enables us to develop it further. Suppose the economy consists of identical individuals consuming a private good $x$ (which may represent a composite aggregate of private goods) whose price is normalized to $p_x=1$, and a public good—say, a swimming pool. The benefits of the swimming pool, $h=h(y, n)$, depend on its size ($y$) and the number of individuals using it ($n$). Let the cost of the service also depend on the size and the number of users, $c=c(y, n)$. Since we start with identical consumers, we shall derive Pareto optimal conditions for the size of the pool and the number of users (the size of the club in Buchanan's paper)—assuming every individual ends up with the same utility. Our problem is therefore:

$$(1) \quad \max u[x, h(y, n)] \quad \text{s.t.} \quad x=I-c(y, n)/n$$

where $I$ denotes individual income. In order to facilitate the use of calculus, $n$ is as-

* Tel-Aviv University and the University of Rochester. This paper was written during my stay in Rochester as a Xerox Visiting Professor. I wish to thank E. Helpman for useful comments.

[1] It will be shown that the Buchanan club solution is efficient, and that the recent criticism of it by Y. K. Ng (1973, 1974) is incorrect.

sumed to be continuous in the analysis which follows.[2]

Notice that the problem defined by equation (1) may have a solution with $n=1$, which means each individual will have his own swimming pool; or even the case where max $u \Rightarrow n=0$, which means that individuals will not consume the services of swimming pools. The theory of clubs assumes that there are commodities for which congestion costs are small or nonexistent when the number of users is small, or that average costs decrease (at least over a range) so that it pays for groups to consume collectively. Furthermore, the theory implies that beyond a certain group size, congestion costs rise faster than average costs fall. Thus, the optimal consuming group $n^*$ is finite and small relative to total population $N$, or $1<n^*<N$.

In order to derive the optimal solution, assuming an internal solution, differentiate equation (1) with respect to $y$ and $n$ to get

$$nu_2h_1/u_1 = c_1 \tag{2}$$

$$c_2 - nu_2h_2/u_1 = c/n \tag{3}$$

Where subscripts denote partial derivatives, or, $u_1 \equiv \partial u(x, h)/\partial x$. Equation (2) is simply the Paul A. Samuelson rule for public goods. It is important to note that this rule should hold independently of the size of the group.[3] Equation (3) implies that if the size of the group is optimal, per capita payment, $c/n$ should equal the marginal social cost (where marginal social cost is defined as the sum of marginal maintenance cost and marginal congestion cost). Using equations (2), (3) and the budget constraint, it is possible to solve for $x^*$, $y^*$ and $n^*$. If $N/n^*=k$ is an integer, the utility of the population is maximized by the formation of $k$ identical clubs.

Ng (1973, 1974) correctly recognizes that by this maximization procedure benefits minus costs per club are not maximized. But he errs by requiring that for Pareto optimality, each club must maximize benefits minus costs. The Ng criterion requires maximization of $nu[x, h(y, n)]$ instead of equation (1). This leads to per capita $u^{**}<u^*$. Therefore, this solution is Pareto inferior. The correct criterion in cost/benefit terms is to maximize total benefits (for the whole community) minus costs; this is consistent with maximization of per capita net benefits. Therefore, the tax-subsidy scheme suggested by Ng is incorrect, and the Pareto efficiency rules derived by Buchanan are the correct ones.

Assuming that $k$ is large enough, and postponing for the moment the possibility that $k \neq$integer, the question arises: Is it necessary that the group will organize as a club? We want to argue that the answer is: no!

Suppose that swimming pool services are organized in an industry which consists of $k$ identical firms. Each firm operates a swimming pool of size $y^*$ and sells $n^*$ season tickets at $P^*_s=c(y^*, n^*)/n^*$ dollars per ticket. We want to show that this industry structure and season ticket price constitute an equilibrium for an appropriately defined decision problem for firms. We need to concentrate only on firms, since it is obvious that under these circumstances individuals will have no incentive to form clubs.

One may think of a firm as an entity which supplies swimming pool services and can control the quality of these services. The quality of these services is a function of the facility size and of the number of users. The $j$th firm is thus free to choose

[2] The following analysis is unaffected if the utility function takes the general form $u=u(x, y, n)$, and/or if $n$ is eliminated from either the utility function or the cost function. In the latter case $\partial(c/n)/\partial n<0$ for every $n$.

[3] In reality, a swimming pool cannot be characterized by one variable such as size. Given equation (2), it is straightforward to generalize the result. Let $y$ denote a vector of characteristics such as size, water temperature, purity of water, etc. Then equation type (2) will hold separately for each element in $y$.

three variables: its facility size $y^j$, the number of users $n^j$ which determines quality and the price of a season ticket $P_s^j$. However, the firm knows that it has to choose $(y^j, n^j, P_s^j)$ so as to assure its buyers of at least the utility level $u^* = u[I - P_s^*, h(y^*, n^*)]$ which is provided by other firms, otherwise it will have no customers. Hence, the firm's decision problem is: Choose $y^j$, $n^j$, $P_s^j$ to maximize

$$(4)\quad P_s^j n^j - c(y^j, n^j) \text{ s.t. } u[I - P_s^j, h(y^j, n^j)] \geq u^*$$

It is easy to see that $(y^*, n^*, P_s^*)$ is a solution to this problem. Hence, the proposed allocation and price are indeed an equilibrium.

Whenever these equilibrium conditions do not hold we have either $P_s \geqq AC$ [and $h \neq h(y^*, n^*)$ when $P_s = AC$] or $P_s < AC$. In the first case it is possible for a new firm to enter the industry and make (an above normal) profit, in the latter case firms will tend to leave the industry.

We have shown that in this case competition will lead to a stable optimal equilibrium. This is the same solution derived for a club equilibrium. Therefore, we can state that Buchanan externalities do not lead to a market failure—and that if there are advantages to organizing into clubs, further explicit assumptions are needed to establish these benefits.[4]

We now turn to the case where $n^*$ is large relative to the market, say $n^* < N < 2n^*$. In this case, any subgroup of $n^*$ people that organizes as a club will be able to maximize its members' utility. Ignoring the problems of the stability of this case, it should be emphasized that this argument in favor of a club is not an outcome of the specific Buchanan model. Suppose instead that we have a case of a private good produced by a U-shaped average cost function. Let the number of consumers that will purchase the total output of the firm, when $P^* = \min AC$, be denoted by $n^*$. Then if $n^* < N < 2n^*$, it pays any $n^*$ to organize a club. The analogy seems incomplete because, in the latter case, members of the "club" will benefit by reselling part of their share in the private commodity to nonmember consumers. This difference is not due to congestion effects but is a result of the nondivisibility assumed in the case of the Buchanan model.

In the next section, we analyze the case where club members can vary the intensity of their use of the club services. When this is done, the analogy between the Buchanan case and the case of firms with U-shaped cost curves becomes complete. Furthermore, the integer problem discussed in detail for the club model by Mark V. Pauly is reduced to the problem of an industry which consists of firms with U-shaped cost curves.

## II

Turning to the intensity of use of club facilities, it is immediately clear that intensity can be affected in various ways. In the case of swimming pools, the club member can change the number of times he visits the place, the length of his stay, and the use of the pool's various services. A member of a professional association is free to choose the activities he takes part in. In the following analysis, however, just one variable represents intensity of use; this will suffice to show the similarity between the club case and the private good case.

When the consumer is allowed to change the number of visits to the swimming pool ($v$), the utility function and the budget constraint are reformulated as follows:

$$(5)\qquad \text{Max } u[x, h(y, v, nv)]$$
$$\text{s.t. } nx + c(y, nv) = nI$$

[4] E.g., additional benefits can arise from: individual preferences with regard to the composition of the club membership, tax advantages and savings in the cost of revenue collection.

We assume that both costs and congestion are functions of the number of visits to the swimming pool ($nv$). The number of visits per consumer ($v$) appears twice in the utility function—once to represent the fact that utility of the individual depends on the number of his visits, and once to represent congestion externalities. Thus, $h_2>0$ $h_3<0$.

The first-order conditions for maximization of equation (5) are given by

$$(6) \qquad nu_2h_1/u_1 = c_1$$

$$(7) \qquad vc_2 - nvu_2h_3/u_1 = c/n$$

$$(8) \qquad u_2h_2/u_1 = c_2 - nu_2h_3/u_1$$

Equation (6) is the Samuelson condition. Equation (7) is similar to equation (3) above (which gives the solution to the optimal number of customers), except that it is corrected for the fact that congestion depends on the number of visits ($nv$). Equation (8) gives the optimal number of visits per customer. Conditions (6)–(8) and the budget constraint in (5) can be solved for optimal $x^*$, $y^*$, $n^*$, $v^*$.

Now suppose that the service is organized on a cooperative basis and that the individual is paying a membership fee of $P_m=c^*/n^*$. His optimal behavior as a club member is then

$$\max u[x, h(y^*, v, n^*v)]$$
$$\text{s.t.} \quad x = I - P_m$$

where he chooses the optimal number of visits without having control over the visits of other members. In this case, his optimal number of visits will be such that $u_2h_2=0$; his marginal utility from visits to the swimming pool will be zero, thus violating equation (8).

Suppose, instead that payment is on a per visit basis such that

$$(9) \qquad P_v = c_2 - nu_2h_3/u_1$$

where all partial derivatives are evaluated at the optimal solution given by equations (6) to (8). The individual budget constraint is now $x+vP_v=I$. Maximizing the utility function, we get $u_2h_2/u_1=P_v=c_2-nu_2h_3/u_1$, which is equation (8). Total revenues of a swimming pool of size $y^*$ that allows $n^*v^*$ visits will be $n^*v^*P_v=n^*v^*(c_2-n^*u_2h_3/u_1)$ —which, by equation (7) is equal to $c^*$.

Thus an optimal-size swimming pool charging a price $P_v$—which by equation (9) is equal to the sum of congestion costs and marginal cost—and which sells $n^*v^*$ tickets per season, will be consistent with Pareto optimality; a club operating on the basis of membership fees will not be optimal. It should be emphasized that a two-part tariff—a combination of a membership fee and a price per visit—is also inefficient. It is easy to show that the same results hold if we assume that demand is not uniformly distributed over time. In this case, it is optimal to have different prices for peak hours and regular hours. These optimal prices, which are of the nature of equation (9), cover total costs. Two-part tariff is still inefficient.

As shown, the club arrangement with membership fees is inefficient. If people prefer to organize clubs and pay membership fees, they must gain other advantages for which they pay in the form of a loss in efficiency.

## III

Until now it has been assumed that all consumers are identical. Introducing two types of individuals, we start by showing that mixed clubs are not optimal.[5] Suppose that this is not the case and that the optimal club consists of $n$ individuals of type one and $m$ individuals of type two. For $y$

[5] Different types for this case require that if segregated clubs of Section I above are formed, each type will prefer a different quality of swimming pool. In the following proof simplifying assumptions are used: (1) people do not vary the number of visits, (2) congestion is a function of the number of members and not of their composition, (3) membership fees are equal for everybody, and (4) the utility of each type of individual is the same in all clubs.

to be optimal, the Samuelson criterion must hold. Or, $nu_2^1h_1^1/u_1^1+mu_2^2h_1^2/u_1^2=c_1$, where superscripts denote the types of individuals, and each type of individual has the utility function of the form of equation (1). Now, rearrange people by forming clubs of homogeneous memberships without changing the sizes of the clubs. Since the number of people in each club does not change, and as long as $y$ does not change, everyone's utility remains unchanged. However, since $u_2^1h_1^1/u_1^1\neq u_2^2h_1^2/u_1^2$, the Samuelson criterion does not hold for the two clubs. Thus, everyone's utility can be increased by adjusting $y$ to the preferences of the now different compositions of the clubs.

This shows that once the population is large enough to justify the existence of many clubs, it is not optimal to have many identical mixed clubs. A Pareto optimal solution is to build two types of clubs by using the procedure of Section I separately for the two types of individuals. (See Pauly and Martin C. McGuire for other proofs of the same proposition.)

In this case, if the service is provided by a competitive market, we shall observe two types of swimming pools—each servicing a population with different tastes or incomes. Furthermore, it is easy to show that this result continues to hold if people can control the number of their visits.

It is interesting to investigate the conditions that lead to a case where people with heterogeneous tastes use the same swimming pool. It seems that in order to allow for this possibility, it is necessary to introduce transportation costs into the model. If consumers do not live in homogeneous neighborhoods, they may prefer a nearby swimming pool not optimally adjusted to their tastes to a distant one which provides preferable services.[6] The competitive solution for the location of firms and their optimal size has not yet been satisfactorily resolved, even under extremely simplified assumptions. (See articles by Harold Hotelling and Lester G. Telser.)

Since we are unable to provide a general solution for this case, we shall restrict ourselves to some partial observations. Suppose that people are identical except for location; suppose further that we wish to maximize utility, subject to the constraint that everybody ends up with the same utility. This implies that the club's membership shares transportation costs. Restricting ourselves to the case where intensity of use is not variable (the Buchanan case presented in Section I of this paper), the club will maximize utility by providing transportation service to its members. Furthermore, it is now necessary to add transportation costs to the cost function. Transportation costs are a function of the location of the club and the number of members. Since the number of members appears in our cost function, the solution for equation (1) covers this case if we include transportation costs in the function. It is clear that for all members it is preferable to select the location of the club so as to minimize transportation costs.[7]

However, it is much more complicated to devise full compensation schemes for heterogeneous populations. Lacking full compensation, the differential utility of customers will be affected by the location of the different facilities; it will pay people to organize a club if they can affect its location. In this case every club has some locational monopoly power. Membership arrangements can be an effective way of eliminating potential competition.

[6] The distribution of consumers in the city is regarded here as exogenous.

[7] If we allow the number of visits per member to vary, then it is optimal that every individual pay a per-visit price equal to his marginal social costs—which are the sum of marginal congestion cost, marginal service costs, and transportation. Equalization of utility can be achieved by having nearby residents pay a positive membership fee which is divided among distant members.

These observations suggest that spatial considerations may play an important role in the economic theory of clubs; further research in this direction is necessary.

## IV

We have shown that where we have both externalities in consumption and increasing returns to scale, market failure need not occur. It seems possible to suggest a generalization of the result: whenever the optimal firm providing a service is small relative to the market and exclusion is possible, the competitive market solution is possible. This formulation does not seem strange on its face; however, we have shown that it covers cases of both externalities and increasing returns to scale. In particular, it covers the Buchanan model—which seems appropriate for many situations.

It has been shown that the Buchanan case is very similar to the case where firms have U-shaped average cost functions. The similarity holds for (1) the case where the optimal firm is small relative to the market, so that competition results in many identical firms and competitive production is Pareto efficient; and (2) the case where the optimal firm is large relative to the market, then increasing returns are dominant and the competitive markets fail.

## REFERENCES

J. M. **Buchanan**, "An Economic Theory of Clubs," *Economica*, Feb. 1965, *32*, 1–14.

H. **Hotelling**, "Stability in Competition," *Econ. J.*, Mar. 1929, *39*, 41–57.

M. **McGuire**, "Group Segregation and Optimal Jurisdictions," *J. Polit. Econ.*, Jan.-Feb. 1974, *82*, 112–32.

Y. K. **Ng**, "The Economic Theory of Clubs: Pareto Optimality Conditions," *Economica*, Aug. 1973, *40*, 291–98.

———, "The Economic Theory of Clubs: Optimal Tax Subsidy," *Economica*, Aug. 1974, *41*, 308–21.

M. **Pauly**, "Cores and Clubs," *Publ. Choice*, Fall 1970, *9*, 53–65.

P. A. **Samuelson**, "The Pure Theory of Public Expenditure," *Rev. Econ. Statist.*, Nov. 1954, *36*, 387–89.

L. G. **Telser**, "On Regulation of Industry: A Note," *J. Polit. Econ.*, Nov. 1954, *77*, 937–52.

INCOME DISTRIBUTION IN DEVELOPING COUNTRIES

# Toward an Adequate Long-Run Model of Income Distribution and Economic Development

By SHERMAN ROBINSON*

This paper has two purposes. The first is to present a strategy for constructing long-run, economy-wide models of developing countries. Second, the paper will outline the components of one possible long-run model which includes the distribution of income. The intent is to provide a modeling framework that can incorporate a number of different types of behavioral submodels in a mutually consistent and interactive manner. The overall framework is explicitly designed to take account of the fact that the component submodels differ widely in their specification of adjustment mechanisms and in their definition of temporal equilibrium.

Any long-run model of a developing country, and especially one which includes the distribution of income, must incorporate a wider range of economic and demographic relationships than is usual in even the largest economy-wide models. It is necessary not only to work with both the income and product accounts, but also to include the interindustry, demographic, and perhaps the household accounts as well. In recent years, Richard Stone and others have been working on developing a system of unified demographic, economic, and social accounts for the United Nations. This work is well advanced and there has been some consideration of the potential of these accounts for broader models of the economy. See Stone (1970, 1973) and United Nations.

The discussion below presents the structure of a model of income distribution and economic development in the long run. The particular model sketched here is designed to include a number of factors which recent empirical work indicates are important. It is based on two different existing models: one developed by Adelman and Robinson and a model developed by a team at the International Labour Office (*ILO*). (See M. J. D. Hopkins, G. B. Rodgers and R. Wéry.) The Adelman-Robinson model is short-run while the *ILO* model is long run. Indeed, the framework presented here can be used to integrate the two different models.

Both the Adelman-Robinson and the *ILO* models yielded similar results which are also largely consistent with other work.[1] Since these results are based on models of two countries—South Korea and the Philippines—they can be thought of as hypotheses which should be tested in other countries and so should influence models designed to study distribution and

* Assistant Professor of Economics, Princeton University. My thanks to Alan Blinder and especially to Irma Adelman for comments on an earlier draft. Remaining errors are my responsibility.

[1] See, for example, Adelman and Cynthia T. Morris and Hollis B. Chenery, et al.

development. Some of the major results are summarized below. The reader who wants a fuller description, including appropriate qualifications, can consult the original studies.

—The studies support the Kuznets U hypothesis that during development the distribution will initially become more unequal and only later become more equal.

—The basic time path of the household distribution is very stable and quite difficult to change by policy intervention.

—The functional distribution is quite unstable and is strongly affected by policy intervention. The functional and size distributions do not necessarily move together.

—The effects of single policies are quickly dissipated. Although the effort required is large, broad based economy-wide policy packages can have significant effects on the distribution. A corollary is that interaction effects and nonlinearities are very important. The overall impact of a policy package is different from the sum of the effects from the separate components.

—Long run or complete adjustment results are often very different from short run or partial equilibrium results.

—In the short run, wage and price adjustments are more important in determining the distribution than are quantity adjustments in both factor and product markets. In the product markets, variations in the agricultural terms of trade have the strongest impact on the distribution.

—Among the quantity adjusting variables, interregional migration, skill upgrading, and education are important in their distributional effects.

These results, along with the general characteristics of developing countries, suggest a number of factors that should be incorporated into any long-run model which includes distribution and also some of the broad design characteristics of such a model.

## I. Desiderata of a Long-Run Model

First, a long-run model should be in the broad tradition of multisector planning models. Existing multisector economy-wide models have, with varying degrees of success and realism, embodied sectoral growth, changes in the structure of production, and changes in the structure of employment.[2] In a long-run model, it would be well to reduce the number of sectors and concentrate instead on incorporating demographic interactions, migration, and changes in the skill structure of the labor force (through education and nonformal training.)

Second, any model which attempts to address distributional questions should endogenously determine both wages and prices. Wage and price endogenous general equilibrium models (*WPE* models) are currently available and can be implemented.[3] In a *WPE* model, the model economy is essentially Walrasian with profit maximizing producers and utility maximizing consumers interacting across product and factor markets. It is a general equilibrium model which solves for wages and prices so that, given nonlinear supply and demand relationships, all product and factor markets clear. The solution problem is that of finding a fixed point for a set of very nonlinear equations in wages and prices.[4]

Third, a distributional model must, in fact, generate the distribution of income in detail. An economy-wide model will naturally generate the functional distribution, at least by broad functional categories such as wages, profits, and rent. An ade-

[2] For a survey, see Charles R. Blitzer, et al.

[3] See Adelman and Robinson, F. Ahmed, K. Dervis, and Lance Taylor and F. Lysy.

[4] For a discussion of solution techniques, see Adelman and Robinson.

quate model, however, should generate the functional distribution in some detail including many income recipient groups and also generate the entire distribution by individual income recipients. A model should also generate the size distribution of income by households and thus explicitly model the grouping of income recipients into households. This is done in the Adelman-Robinson model. For related techniques, see Hopkins, et al., and Erik Thorbecke, and Jati Sengupta.

Finally, a question that must be dealt with in formulating a long-run model is how to incorporate the fact that different parts of the economic system adjust at very different rates. There is an explicitly dynamic problem in defining the nature of equilibria in a model which seeks to describe a developing country over an extended period of time. A long-run model must deal with two different kinds of adjustment behavior. First, there are markets such as product markets in which prices and quantities adjust quickly and excess demands are reduced to zero within a relatively short period. Second, there are markets in which excess demands or supplies can persist for an extended period. For example, continuing technological change in industry and continuous lags in investment and migration responses may cause marginal products in the urban sector to be above those in the rural sector for a long period of time.[5] In a long-run model, it is thus necessary to model the disequilibrium behavior of these markets and so explicitly model the laws of motion of the dynamic economy.

The disequilibrium dynamics required of a long-run model can best be represented by quantity adjustment models (following Janos Kornai). Short period disequilibrium reflects both price and quantity disequilibria and, in the usual analysis, prices will adjust to clear the market by moving factors along their supply and demand curves. One can assume that factor immobility will persist in the short run and will result in differential wages and rentals among sectors. In the long run, one must model the quantity response to the disequilibrium which is reflected in the short-run differential wages and rentals. The experience of many economies suggests that full equilibrium is never reached; i.e., that wages and rentals are not equated across sectors and regions, even over long time periods.

## II. Structure of a Long-Run Model

This section outlines a proposed structure for a long-run model. The structure proposed here is intended to encompass a number of different interacting component models in a unified framework.

The model framework is presented schematically in Figure 1. In each period, the solution of the overall model is divided into two stages. In the particular models diagrammed here, the Stage 1 model is a *WPE* general equilibrium model whose solution gives as outputs: production, employment, wages, prices, and the distribution of income. The *WPE* model framework is flexible enough to accommodate a wide variety of behavioral assumptions and institutional constraints. The Stage 2 model consists of a number of sub-models which take as inputs different parts of the Stage 1 solution and yield as output the quantity response of factor supplies and all other exogenous variables needed for Stage 1 in the next period. Six different sub-models are listed in Figure 1 with varying degrees of linkage to the Stage 1 solution and among themselves.

In order to understand the division between Stages 1 and 2, there are three notions of equilibrium which need to be dis-

[5] See M. Bruno and Robinson for examples of two-sector disequilibrium growth models.

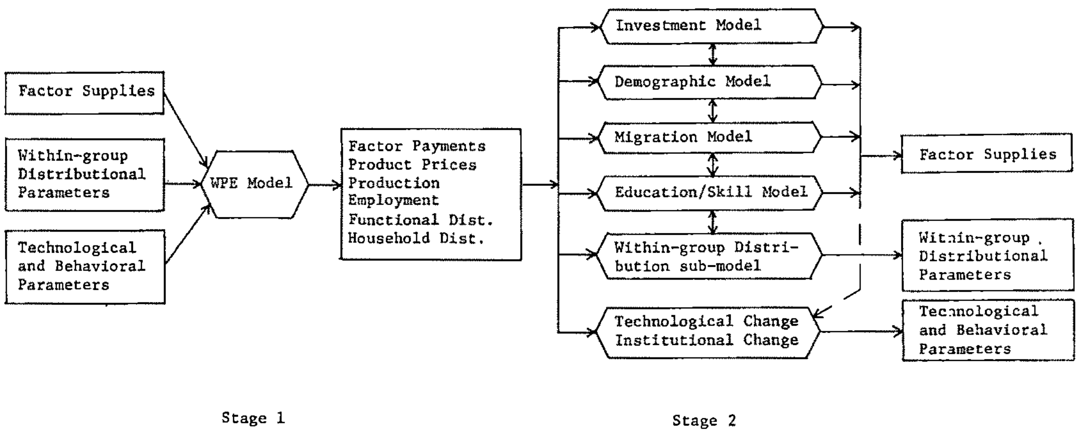



FIGURE 1

tinguished. First, in Stage 1, short-period equilibrium is defined as zero excess demand in product and factor markets subject to a number of constraints on factor mobility. The operation of these constraints leads to sectoral differentials in wages and rates of return to capital in place. The elimination of such differentials defines a second notion of equilibrium, that of full neoclassical equilibrium. Full neoclassical equilibrium is not attained in Stage 1; indeed, the solution of Stage 1 describes the degree of factor price disequilibrium. The third notion of equilibrium is that of intertemporal or long-run equilibrium in which all expectations are validated and in which the first two types of equilibrium prevail continually. The Stage 2 model consists of a number of different quantity adjustment submodels which, when solved together, yield the response of factor supplies to the situation described by the solution of Stage 1. The solution of the two stages yields neither full neoclassical equilibrium nor full intertemporal equilibrium, although the Stage 2 quantity adjustment submodels should move in an equilibrating direction.

The distinction between Stages 1 and 2 is really one of different adjustment rates. In Stage 1, factor supplies are assumed fixed, where "fixed" may have a number of different meanings. For example, one might assume that labor and capital are completely immobile among sectors, or one might assume that they are partially mobile. One might also assume that some wages are fixed exogenously in the tradition of two-sector excess labor models. (See Dervis and Ahmed.) The Stage 2 submodels together describe the dynamic response of the economy. Their specifications will clearly depend on the nature of the constraints imposed in Stage 1. For example, the migration model would be very closely related to the Stage 1 model, with migration being a function of regional income differentials. The investment model requires two components, one to determine aggregate investment (and savings) and one to determine the sectoral and regional allocation of investment, which should clearly respond to the differential rates of return solved in Stage 1. The challenge is to model the degree of response.

The educational, demographic, and migration submodels are shown as being interrelated. The *ILO* model is perhaps the most satisfactory current model in including the various interrelationships. Stone (1970, 1973) and Frederic Shorter are

examples of alternative models. Of the three submodels, the migration model would be the most sensitive to the solution of Stage 1. A general framework for incorporating the interrelationships among the submodels is suggested in Robinson and Dervis.

The technological and institutional change submodels are shown as interconnected with the other submodels. Given the current state of knowledge, this represents more wishful thinking than a reflection of existing models. The roles of technological and institutional change in growth—especially as they differ by sectors and regions—are among the least well understood economic processes and among the most important.

There are a number of different ways in which the distribution of income can enter the model stages. Following Adelman-Robinson, it is possible to have the Stage 1 model solve for the mean incomes of a number of different functional groups and generate the overall distribution by assuming known frequency distributions for the within-group distributions whose parameters are specified exogenously to Stage 1. In the diagram, these would be the exogenous distributional parameters shown feeding into Stage 1. The overall distribution is calculated by simply summing the intergroup distributions whose parameters are known. A Stage 2 submodel would update the exogenous distributional parameters, either independently or as a function of the Stage 1 solution. How the within-group distributions vary over the long run is not well understood either theoretically or empirically and represents a potentially fruitful area for further research. One possibility, for example, is to argue that the within-group distributions are a function of the age and education structure within a given skill or functional category. Thus changes in the within-group variance of income might be specified as a function of changes in the within-group age and educational structures which are generated by the educational and demographic submodels.

## III. Conclusion

There are currently no long-run models that fit into the framework presented here. The Adelman-Robinson model is explicitly a short-run model but does incorporate the basic idea of modelling stages presented here. Being a short-run model, it concentrates in Stage 2 on the investment and migration submodels and incorporates a number of financial and monetary effects. The *ILO* long-run model incorporates many of the Stage 2 submodels but has a much simpler Stage 1 (not a *WPE* model). There are also *WPE* models which incorporate some of the Stage 2 components directly into the Stage 1 models, remaining within an explicitly equilibrium framework. For example, see Dervis and Ahmed.

The modelling framework proposed can be used both for building new models and for classifying existing ones. For example, within this framework, the J.R. Harris-M.P. Todaro migration model can be seen as an attempt to view migration as an equilibrium process that can be modeled in Stage 1 rather than as a disequilibrium adjustment that is better modeled in Stage 2.

The separation of market clearing and quantity adjusting dynamic models into two distinct stages should permit much more realistic modeling of dynamic processes. The distinction between equilibrium and disequilibrium models has been widely recognized and is currently being used by those who work with Keynesian macro-models (who, however, consider a single set of markets and compare short-run disequilibrium with long-run equilibrium). With the development of *WPE*

general equilibrium models, builders of planning models can now exploit the notion and, at the same time, expand the boundaries of their models.

## REFERENCES


I. Adelman and C. T. Morris, *Economic Growth and Social Equity in Developing Countries*. Stanford 1973.

——— and S. Robinson, "A Wage and Price Endogenous General Equilibrium Model of a Developing Country: Factors Affecting the Distribution of Income in the Short Run," Washington, International Bank for Reconstruction and Development, Development Research Center, Sept. 1975.

F. Ahmed, "Migration and Employment in a Multisector Model: An Application to Bangladesh." Princeton Univ., unpublished Ph.D. dissertation, 1974.

C. R. Blitzer, P. B. Clark, and L. Taylor, *Economy-Wide Models and Development Planning*, Oxford 1975.

M. Bruno, "Estimation of Factor Contribution to Growth Under Structural Disequilibrium," *Intern. Econ. Rev.*, Feb. 1968, *9*, 49–62.

H. Chenery, M. Ahluwalia, C. L. G. Bell, J. Duloy, and R. Jolly, *Redistribution with Growth*, Oxford 1974.

K. Dervis, "Planning Capital-Labor Substitution and Intertemporal Equilibrium with a Non-linear Multi-sector Growth Model," *European Econ. Rev.*, Jan. 1975, *6*, No. 1.

J. R. Harris and M. P. Todaro, "Migration, Unemployment and Development: A Two Sector Analysis," *Amer. Econ. Rev.*, March 1970, *60*, 126–142.

M. J. D. Hopkins, G. B. Rodgers, and R. Wéry, "A Structural Overview of Bachue-Philippines," Geneva, International Labour Organization, World Employment Program, Population and Employment Working Paper No. 20, May 1975.

J. Kornai, *Anti-Equilibrium*, Amsterdam 1971.

S. Robinson, "Sources of Growth in Less Developed Countries: A Cross-Section Study," *Quart. J. of Econ.*, Aug. 1971, *85*, 391–408.

——— and K. Dervis, "Income Distribution and Socio-economic Mobility: An Approach to Distribution Planning." Princeton, Research Program in Development Studies, Discussion Paper No. 49, Sept. 1974.

F. C. Shorter, *Computational Methods for Population Projections: With Particular Reference to Development Planning*, New York 1974.

R. Stone, "Economic and Demographic Accounts and the Distribution of Income," *Acta Oeconomica, 11* (2–3), 1973, 165–179.

———, *Mathematical Models of the Economy and Other Essays*. London 1970.

L. Taylor and F. Lysy, "Brazilian Income Distribution, 1960–1970: Confronting the Conflicting Interpretations," Paper presented to the Econometric Society, Third World Congress, Toronto, Aug. 1975.

E. Thorbecke and J. K. Sengupta, "A Consistency Framework for Employment, Output and Income Distribution: Projections Applied to Colombia," Development Research Center, International Bank for Reconstruction and Development, 1972.

United Nations, Department of Economic and Social Affairs, *Towards a System of Social and Demographic Statistics*, United Nations, ST/ESA/STAT/SER.F/18, New York 1975.

# Income Distribution and Development: Some Stylized Facts

By MONTEK S. AHLUWALIA*

In recent years, the relationship between income distribution and the process of development has come under increasing scrutiny. Much of the debate has focused on the hypothesis, originally advanced by Simon Kuznets, that the secular behavior of inequality follows an inverted U-shaped pattern with inequality first increasing and then decreasing with development. This hypothesis has become so much a part of the conventional wisdom on this subject that it has generated considerable skepticism about the welfare implications of the development process. Indeed, on some interpretations, developing countries face the grim prospect not just of increasing relative inequality, but also of declining absolute incomes for the lower income groups.

The object of this paper is to reexamine the empirical basis for this hypothesis, using a recent compilation of cross-country data.[1] We have used multiple regression to estimate cross-country relationships between inequality, as reflected in the income shares of various percentile groups, and selected explanatory variables reflecting different aspects of the development process. Needless to say, associational relationships of this type do not establish causality. They are better described as "stylized facts," which can be observed, but which then need to be explained by an appropriate theory.

## I. Short and Long Term Relationships

As a first step, it is useful to distinguish between two types of relationships that might obtain between income distribution and development; there may be a long-term or secular relationship between the degree of inequality and the level of development, and there may be a short-term relationship between inequality and the rate of growth. Both types have been discussed in the literature, but the importance of distinguishing between them has not been adequately recognized.

Kuznets' hypothesis clearly referred to a secular relationship generated by long-term changes in economic structure. It was, after all, based on an observed trend in this century towards a reduction in inequality in some of the now developed countries, which Kuznets contrasted with the "more probable" trend of widening inequality in the early stages of their industrialization. In contrast with this type of long-term relationship, some of the recent debate has focused on much more short-term phenomena. Thus it has been suggested that high growth rates in some less developed countries (*LDC*), observed over comparatively short periods, e.g., Brazil between 1960 and 1970, have led to a marked increase in income inequality. While such an outcome can be explained simply in terms of Kuznets' hypothesis—a higher rate of growth raises

* Development Research Center, World Bank. I am indebted to Marienne Lehwing for computational assistance. The views expressed in this paper are those of the author and not of the World Bank.

[1] Income distribution data for 62 countries (including 14 developed countries and 6 Socialist countries) were taken from a recent compilation of available data (see Shail Jain). Data on explanatory variables for years corresponding to the observations on income distribution are taken from the World Bank's data files.

the level of development, which in turn affects inequality—it is noteworthy that the debate on this issue has a somewhat different flavor. There is a distinct suspicion that there are short-term forces operating independently of the long-term phenomenon which generate higher inequality as a consequence of faster growth. For example, if growth is concentrated in particular regions or sectors (as is very likely to be the case), then lags in labor mobility may create factor market disequilibria which generate significant income differentials. It is easily seen that such differentials need to be distinguished from income differentials reflecting long-term structural factors such as the scarcity of labor skills. They are superimposed on the structurally determined income differentials and may be seen (and are often defended) as necessary lubricants to overcome some of the frictional resistance to the pursuit of high growth rates.

We can test for each of these relationships by appropriate choice of explanatory variables in our regression equations. The secular relationship can be tested by including per capita *GNP* (in various transformations) as an explanatory variable and the short-term relationship can be tested by including the rate of growth of *GDP* over the past ten years as an additional explanatory variable. The relevant findings of our cross section analysis can be summarized as follows. See equations (1a), (2a), (3a) and (4a) in Table 1.

(1) There is substantial confirmation of a statistically significant relationship between the income shares of various percentile groups and the logarithm of per capita *GNP*. Furthermore, the relationship is clearly not monotonic since a significant relationship is identified only when the per capita *GNP* variable is entered in quadratic form. This formulation generates the familiar U-shaped pattern: the income share of the lower income groups (alternatively defined as the lowest 60 percent and the lowest 40 percent) first declines and then rises with increasing per capita *GNP*, while the share of the upper income group (top 20 percent) describes the opposite pattern. Similar results have been reported by Irma Adelman and Cynthia Morris and Hollis Chenery and M. Syrquin.

(2) By contrast, there is no evidence of an independent short-term relationship between the level of inequality and the rate of growth of *GDP* over the recent past. The coefficient on this variable was insignificant in all the equations estimated in experimenting with alternative functional forms and different combinations of other explanatory variables.

The lack of a significant relationship between inequality and the rate of growth is potentially important in defining the nature of the relationship between distribution and development. It suggests that while there may be a secular time path for inequality which developing countries must traverse and which contains a phase of increasing inequality, there is at least no evidence that faster growing countries show higher inequality at the same level of development than slower growing countries. If this is true, policymakers are perhaps best advised to think of the rate of growth as determining essentially the speed of transition through the different phases of development and inequality: higher growth rates accelerate the transition without necessarily generating greater inequality than can be structurally expected in each phase.

We hasten to add that our cross-section results provide only weak evidence for this point of view. For example, we cannot deny that particular types of growth

processes may have markedly adverse effects on inequality. The cross-section evidence only establishes that fast growers do not systematically display any such pattern i.e., if there are countries in which this is true, there are others whose experience is of the opposite variety. Recognizing this diversity of experience is perhaps the most important lesson to be learned from the data. At the very least, it shifts the focus of debate away from a naive (and quite possibly dangerous) suspicion of high growth rates as such towards an examination of the particular nature of growth in different countries and the implications of different types of growth for inequality. Inquiries along these lines must figure high on any agenda of research on income distribution, but until their results persuade us to the contrary, the verdict on growth rates need not be unequivocally pessimistic.

## II. The Secular Relationship

We now turn to a more detailed examination of the long-term or secular relationship identified above. The observed U-shaped relationship between inequality and per capita *GNP* obviously reflects the net effect on inequality of a number of structural changes occurring with development. The nature of these changes has been extensively discussed in the literature, albeit only in general terms.[2] One of the mechanisms generating an increase in relative inequality in the early stages of development is the shift of population from the low income, slow growing, traditional sectors to the high income, faster growing, modern sectors of the economy. The difference in mean incomes between sectors suffices for such a process to produce a phase of increasing inequality, and this tendency is further reinforced if the modern sectors are also characterized by greater within-sector inequality. There are other factors which operate to reverse the process of increasing inequality in the later stages of development. For one thing, as the modern sector expands, it absorbs larger proportions of the labor force into high income employment, thus reducing the pressure of population in the traditional sectors and thereby narrowing intersector income differentials. Equally important, there are long-term forces which operate to reduce inequality within the modern sector. The cumulative impact of an expanded education system and a long established modern sector is to create a highly trained labor force with a more equal dispersion of skills which generates both an increase in the share of wage income as well as greater equality in its distribution. This tendency is further strengthened by improvement in labor organization.

We have attempted to quantify the impact of some of these mechanisms by including in our regression equations selected explanatory variables which reflect these structural changes. In general, we find that explanatory variables reflecting expansion in education, diminution of demographic pressures, and changes in the structure of production in favor of the modern sector are significantly related to the observed cross country patterns in inequality. The inclusion of these variables, in addition to the quadratic in per capita *GNP*, substantially improves the percentage of variation explained by the regression equations. See Table 1, equations (1b), (2b), (3b) and (4b). Our findings can be summarized as follows:

(1) Improvements in the quality of human resources have long been thought to be a major force for reduction in inequality. (See especially, Adelman and Morris). Measuring improve-

[2] Systematic historical studies of their impact in particular countries have yet to be conducted—the lack of time series data being a major limitation.

TABLE 1—CROSS-COUNTRY REGRESSIONS EXPLAINING INCOME SHARES
(Values in Parentheses Are T Ratios)[a]

| | Dependent Variable: Percentage Income Shares | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Top 20 Percent | | Middle 40 Percent | | Lowest 60 Percent | | Lowest 40 Percent | |
| Explanatory Variables | Eq. (1a) | Eq. (1b) | Eq. (2a) | Eq. (2b) | Eq. (3a) | Eq. (3b) | Eq. (4a) | Eq. (4b) |
| 1. Constant | −65.27 | −9.07 | 89.47 | 31.15 | 128.60 | 110.20 | 75.77 | 77.93 |
| | (2.28) | (0.27) | (4.68) | (1.34) | (5.95) | (4.14) | (5.47) | (4.11) |
| 2. *log* per capita *GNP* | 96.94 | 50.35 | −48.21 | −3.07 | −81.39 | −62.66 | −48.70 | −47.28 |
| | (4.47) | (2.13) | (3.33) | (0.19) | (4.98) | (3.40) | (4.65) | (3.60) |
| 3. [*log* per capita *GNP*]$^2$ | −18.86 | −8.16 | 9.76 | 0.52 | 15.48 | 10.14 | 9.09 | 7.65 |
| | (4.85) | (1.98) | (3.76) | (0.19) | (5.28) | (3.16) | (4.84) | (3.35) |
| 4. Growth rate of *GDP* | −0.22 | −0.11 | 0.13 | 0.001 | 0.19 | 0.18 | 0.08 | 0.11 |
| | (0.48) | (0.32) | (0.44) | (0.01) | (0.54) | (0.63) | (0.37) | (0.55) |
| 5. Literacy rate | | −0.09 | | .03 | | 0.09 | | 0.06 |
| | | (2.21) | | (1.16) | | (2.87) | | (2.56) |
| 6. Secondary school enrollment | | −0.14 | | 0.12 | | 0.07 | | 0.02 |
| | | (2.48) | | (3.02) | | (1.6) | | (0.74) |
| 7. Growth rate of population | | 3.59 | | −2.40 | | −2.54 | | −1.19 |
| | | (4.29) | | (4.20) | | (3.89) | | (2.56) |
| 8. Share of agriculture in *GDP* | | −0.25 | | 0.21 | | 0.13 | | 0.04 |
| | | (2.23) | | (2.70) | | (1.43) | | (0.65) |
| 9. Share of urban population | | −0.10 | | 0.04 | | 0.08 | | 0.06 |
| | | (1.68) | | (1.00) | | (1.82) | | (1.79) |
| 10. Dummy for socialist countries | −20.27 | −9.41 | 8.14 | 0.85 | 17.76 | 10.24 | 12.13 | 8.57 |
| | (6.72) | (3.27) | (4.04) | (0.40) | (7.80) | (4.56) | (8.32) | (5.35) |
| $\bar{R}^2$ | .58 | .76 | .46 | .68 | .60 | .76 | .59 | .69 |
| *F* | 21.6 | 22.3 | 14.1 | 15.5 | 24.16 | 22.4 | 22.86 | 16.21 |
| *SEE* | 6.4 | 4.6 | 4.3 | 3.2 | 4.9 | 3.6 | 3.11 | 2.6 |
| Estimated turning point of quadratic (per capita *GNP* in US$) | 372 | 1217 | 294 | — | 425 | 1230 | 477 | 1231 |

[a] For our sample, a T value of 1.68 indicates significance at the 10 percent level for a two-tailed test.

ments in this dimension obviously poses formidable conceptual and empirical problems. At the very least, we need data on the skill structure of the labor force, but even this is not available for most of the countries in our sample. We have therefore experimented with the literacy rate and various definitions of the school enrollment rates as explanatory variables roughly reflecting the quality of human resources. As shown in Table 1, improvements in literacy have a beneficial impact on the income share of the lowest 40 percent while the secondary school enrollment rate has a beneficial impact on the income share of the middle 40 percent, probably reflecting the exclusion of the lower income groups from access to schooling. It is important to note that in each case, the increase in income share occurs at the expense of a reduction in income share of the top 20 percent—an unambiguous improvement from a welfare point of view. We note in passing that in the case of the middle income groups the inclusion of education as an explanatory variable has the effect of making the per capita *GNP* variables insignificant.

(2) There are strong a priori reasons for supposing that demographic pres-

sures have an important impact on income inequality. High growth rates of population are likely to generate greater inequality by perpetuating the phenomenon of "surplus labor," in the sense that a large proportion of the work force remains locked into low income employment in the traditional or informal sectors of the economy. The cross-section results provide substantial support for this point of view. The rate of growth of population is highly significant as an explanatory variable in all our equations, and as we would expect, it has an adverse effect on the income shares of the lower and middle groups while raising the income share of the top 20 percent.

(3) We find some support for Kuznets' original suggestion that changes in the structure of production provide a mechanism through which development affects inequality. The process Kuznets had in mind has two aspects —declining relative importance of agricultural production in total output, and a shift in population from the low income agricultural sector to the high income modern sector. We have modeled each aspect separately by using two explanatory variables, the share of agriculture in total *GDP* and the share of urban population in the total population. We find that each variable is significantly related to the pattern of inequality but their effects on the income shares of different groups are not identical. The share of agriculture in *GDP* is not significantly related to the income shares of the lowest groups, but it is positively related to the income shares of the middle groups and negatively related to the income shares of the top 20 percent [equations (2b) and (1b)]. By contrast, the share of the urban population in the total has no effect on the income share of the middle group, but it is positively associated with the income shares of the lowest groups and negatively with the income share of the top 20 percent [equations (3b), (4b) and (1b)]. This suggests an interesting asymmetry in the distributional impact of structural change. As the share of agriculture in *GDP* declines with development, there is a relative shift of income away from the middle and towards the upper groups. Alongside this process, however, development also generates a shift of population to the modern or urban sectors, and this process, appears to favor the lowest income groups at the expense of the rich. This asymmetry between the impact on the middle and lower groups is not entirely implausible. At some risk of stretching our evidence too far, we can speculate that the decline in relative importance of agriculture is likely to have its strongest impact on small and middle sized landholding cultivators who may dominate the middle income groups, while the parallel process of increasing urbanization favors the low income population because it reflects increased employment in urban areas and a reduced pressure of population in the rural areas.

(4) Finally, we find that socialist countries display markedly greater equality than others as shown by the positive and significant coefficients on the dummy variables in equations (3a), (3b), (4a) and (4b).

These findings provide some clues to the casual mechanisms that lie behind the conventionally described U-shaped behavior of inequality. They are no more than clues because the essential complexity of a dynamic process cannot be

captured by the variables used in our analysis, which appear to produce a worsening of relative inequality in the early stages of development.

### III. Relative Inequality and Absolute Impoverishment

This prospect of declining income shares of the lower income groups raises obvious questions about the welfare aspects of development. In pursuing these questions, we should distinguish between two alternative views of the underlying process. The pessimistic view holds that this pattern reflects a process of absolute impoverishment of lower income groups in developing countries. (See for example Adelman and Morris.) Such an outcome cannot be ruled out a priori. It may result from an erosion of traditional economic structures against the impact of an aggressively expanding modern sector, competing against traditional sectors both for markets and resources. The resulting disruption of the preexisting economic system could lead to absolute impoverishment for some groups. There is, however, another view of the process that is somewhat less pessimistic. On this view, the worsening in relative inequality occurs not because there is a decline in the absolute incomes for the lower income groups but because rates of growth of income are lower than for upper income groups. Thus if development occurs through opportunities for economic expansion becoming available initially to upper income groups, and if these groups have relatively weak income linkages with lower income groups, we would expect development to be accompanied by an increase in relative inequality, although no absolute impoverishment need occur. The essential distinction between these two views lies in whether we assume that the disruption of low income traditional economic activities is an essential precondition for growth of the modern sector, or whether the problem is primarily one of insufficient linkage. In the latter case the problem appears somehow more tractable since we are no longer dealing with an inherent contradiction in the development process.

Cross-section analysis cannot take us very far in choosing between two such opposite views of the development process; the choice raises issues which are central to our understanding of the nature of development and these issues call for much fuller theoretical and empirical investigation. But, in keeping with our stated objective, we can at least document the stylized facts of cross-country experience. A crude measure of the average absolute income level in each percentile group can be obtained by multiplying the ratio of income shares to population shares of the group by per capita *GNP* of the economy (data on per capita personal income, which is the correct variable to use, are not available). Using equations (3a), 3b), (4a) and (4b) to estimate income shares of the lowest 60 percent and lowest 40 percent at different levels of per capita *GNP*, we can obtain estimates of absolute income levels in these groups at different levels of development. None of these estimates shows any decline in the absolute average income of the poor as per capita *GNP* rises. Alternatively, we have used the actual income share data to estimate the average absolute income of the lowest 40 percent and the lowest 60 percent. This estimate of average absolute incomes of the poor was then used as a dependent variable (in logarithmic form) in cross-country regression equations. Testing for a quadratic relationship with the logarithm of per capita *GNP* for the economy as a whole, we find that the negative term in the quadratic is not significant. This suggests that while there is substantial support for the hypothesis that relative inequality increases in the early stages of

development, the cross-country evidence does not support the stronger hypothesis that there is also an actual decline in the absolute income level of the poorer groups in this phase.

We should be careful not to read too much into this finding. Clearly, the percentile groups are very large and the observed behavior of average incomes for such groups may hide significant declines in the absolute incomes of particular socioeconomic groups which may be offset by above-average growth in income for other groups. Besides, the relationships discussed above are at best descriptions of "average" cross-country behavior; leaving ample room for intercountry differences. Nevertheless, it is important to record that the stronger hypothesis of declining absolute incomes for large sections of the population is not so unequivocally established by cross-country data as to be uncritically accepted as one of the "stylized facts" about development.

## REFERENCES

I. Adelman and C. T. Morris, "Economic Growth and Social Equity in Developing Countries," Stanford 1973.

H. Chenery and M. Syrquin, "Patterns of Development 1950–1970," Oxford 1975.

S. Jain, "The Size Distribution of Income: A Compilation of Data," World Bank Staff Working pap. no. 190, Washington 1975.

S. Kuznets, "Economic Growth and Income Inequality," *Amer. Econ. Rev.*, Mar. 1955, *65*, 1–28.

# Uneven Development and Dependent Market Economies

*By* Michael H. Best*

This paper addresses the issue of inequality from a historical perspective and is organized around three themes. First, a country's factor endowments are not determined exogenously but by powerful systemic forces, one of which is the nature of a country's integration into the world economy. Second, the class structure in the periphery mediates between the imperatives of accumulation in the center and factor endowments in the periphery. Third, income distribution in the periphery is tied to the international division of labor and the corresponding patterns of commodity specialization. Based on the analysis presented, implications for inequality-reducing strategies within poor countries are explored. The basic conclusion is unequivocal: a sharp reduction in inequality requires far-reaching structural changes in the relations of production.

## I. Growth in the Center and the Formation of Periphery Endowments

The conditions for rapid and sustained capital accumulation in Western Europe were forged by the historical conjuncture of three interdependent factors: a propertyless labor supply, a capitalist class or state in command of the productive forces and favorable economic relationships with militarily weaker regions of the world. Interactions with other regions provided important elements in the structural foundations of expansion in West Europe. The spread of commerce undermined the social relations that governed the feudal mode of production and stimulated the market relations of commodity production; in the process population moved from the land to the burgeoning capitalist enterprises. Expansionary forces within the centers created pressures for the opening of new territories for precious metals and other cheap or strategic raw materials, for investment opportunities and for market outlets. National rivalries, in part engendered by this expansion, in turn intensified it.

The crucial factor in establishing commodity production, a cheap labor supply, was recruited in the hinterland by a massive displacement of population both within and between hinterland regions. Besides the ubiquitous migrations that accompanied the penetration of market oriented production within the hinterland regions, many of the dislocated people were removed from their continental shores including 30 to 40 million Africans transported to the Americas; indentured Indian labor shipped en masse to Malaya and Ceylon, and in lesser amounts to Trinidad, British Guiana, Jamaica, East Africa, Mauritius and Fiji; and Chinese labor imported to mine gold in South Africa and to build railroads on the West Coast of North America. A variety of labor recruitment mechanisms ensured labor force dependence upon landlords and mine owners including slavery, servitude, share cropping, debt peonage, vagrancy laws, and wage-labor itself. The labor reserve was

* Associate Professor of Economics, University of Massachusetts, Amherst. I wish to thank Jane Humphries and Stephen Resnick for discussion of these and related topics and Nancy Folbre and Rick Wolff for comments.

replenished by ruinous competition from metropolitan substitutes for traditional products, the extensive control of available land by upper class groups which limited landholding opportunities for the working population, and, as will be discussed later, by high birthrates.

The integration of the hinterland into the merging world economy did not, as expected, induce capital accumulation in the hinterland. While the expansion of markets did create the potential for hinterland capitalist development, it also generated a class structure that blocked hinterland accumulation. Whereas in the center production was dominated by an industrially based bourgeoisie, in the hinterland the forces of production were controlled by a derivative owning class composed of planters, merchants, and mineowners. None of these groups became an autonomous force propelling accumulation. Instead they were interlocked with and subservient to the owning classes of the center; their fortunes were dependent not upon the vitality of the domestic economy but upon the pace of center expansion. Once primary product specialization was established, free trade and entrenched domestic interest groups were often sufficient to block the diversification of hinterland economies; in other cases mercantilist tariff policies and military force were required. Furthermore, the widespread evolution of nonwage-labor systems of labor recruitment impeded the functioning of capitalist labor markets and contributed to a persistent unevenness in productivity across regions, sectors and enterprises.

In short, the specific forms of world specialization and trade that emerged were caused less by "natural" endowments than by an international political context in which the dynamics of capitalist accumulation governed product choice and imposed factor endowments on the hinterland. The mechanisms which established product specialization and labor recruitment varied from region to region, but the consequences were similar: an international division of labor was created in which the hinterland specialized on primary products and the center on industrial commodities.

## II. Class Structure, Commodity Specialization and Income Distribution

The metropole/hinterland class differences are reflected in and reproduced by the historical dichotomy in patterns of commodity specialization, patterns that set into motion forces with very different implications for accumulation and income distribution. Specialization in industrial commodities and capital goods opened possibilities for product diversification, enhanced the introduction of new technologies and product cycles which imparted a dynamic expansionary thrust to the economy, and generated a level of labor income sufficient to stimulate the production of mass produced industrial commodities.

The dynamics of primary product specialization imposed a different experience on the hinterland. Such production, comprising an isolated segment of a center's productive system meant that in the hinterland linkages to other production stages and diversification to other product lines were limited. Concentration on cash crop exports made the region increasingly dependent upon foreign sources not only for manufactured commodities, capital, and invisibles but basic foods as well. In addition, the landowning class was not driven to accumulate capital and innovate by competitive pressure and the threat of technological obsolescence and financial ruin as their industrial brethren of the centers. Technological advancements were thwarted also on the minifundia by the small scale of production.

Primary commodity specialization and the hinterland class structure tended to

cumulatively reinforce each other. The dependence of labor on planters and mineowners for employment generated an extremely unequal distribution of income with a working class market too small to stimulate local production of manufactured goods, yet a large upper class demand for luxury imports which increased the dependence upon traditional exports. The limited employment opportunities in manufacturing ensured the dependence of the labor force upon planters and mineowners and depressed labor costs.

The post-World War II period of spreading political independence and nationalist movements ushered in a series of technician type, progress-oriented governments that systematically sought to generate an environment by which landed and merchant surplus could be transformed to industrial capital through a variety of trade, fiscal, and monetary measures. However, given the class configurations, international environment and capital intensive bias of imported technologies, the process created artificial industrial monopoly capital which was soon stalled by the chronic impediment to capitalist expansion: the tendency for productive capacity to overshoot effective demand.

The historic remedies to realization problems are less appropriate for today's poor countries. Expanding foreign sales are made difficult by the intense competition for exports amongst advanced capitalist countries all burdened by excess productive capacity, by the oligopolized and partitioned marketing channels dominated by the multinational corporations, and by protectionist government tariff and subsidy policies within the industrialized centers. Expanding internal sales by Keynesian income redistribution measures, given an elastic demand for a single country's primary product exports, would likely induce a foreign exchange crisis by driving down the sales of traditional exports both from higher labor costs and expanded domestic consumption and by increasing the demand for imports.

The post-World War II period has witnessed a shift in the locus of control toward the multinational corporations. Given their near monopoly over finance capital and technology the multinationals, without mediation, determine employment, product choice, investment, and pricing decisions and directly extract surplus through transfer pricing and fees for technology, licenses and other services. Furthermore, as recent events in Chile illustrate, when political order and the historic class configurations are threatened, owners of foreign capital continue to work in alliance with local and home government police and military agencies to maintain the status quo.

This period has also seen the continuation of another pressure from the center, mediated through the hinterland class structure, on factor endowments in the hinterland. During the process of industrialization, center countries experienced a strikingly similar demographic transition. Death and birth rates dropped from a position of balance of roughly 40–50 per thousand to a position of near balance in which a decline in birth rates to 15–20 per thousand followed within a few generations an earlier decline in death rates to 10–15 per thousand.

In the primary producing areas one side of this demographic transition took place. The need for labor in the hinterland contributed to the transfer and development of medical techniques designed to cut death rates to levels comparable with the centers. But a drop in birth rates never occurred. Instead, the commonly cited explosion in hinterland population levels has contributed to the increasing per capita income gap between rich and poor countries.

Birth rates have not dropped for good

reason. Given the class structure the only avenue for lower class people to improve family income and obtain security in old age is to have a relatively large family whose members can either work land and generate savings for the family or emigrate to the city and send back remittances which can then be used to acquire additional land and economic security. A social system based, instead, on cooperative ownership of property and social appropriation of surplus does not place on the family alone the burden of generating an income stream, meeting the costs of an education, providing social security for old age, and saving for unforeseen health needs or economic advancement. Creating social institutions to share their burdens takes pressure off the family. Alternatively in a private market system a family is forced to respond to social problems which originate in the sphere of production with individualistic adjustments in personal behavior. One of the few private responses available to people in poor market countries is the size of the family. The relative ineffectiveness of numerous progeny as a remedy for social problems implies less about sexual drives, time horizons, or intelligence of the people than about the limited responses available for dealing with such problems under private market systems.

## III. Inequality Reducing Structural Reforms

The preceding has two basic implitions for social change. First, the path followed by now rich countries is not open to contemporary poor countries. Most are not likely to benefit by the international transfer of surplus or enhancement of conditions for accumulation from a hinterland weak in military and strong on raw materials as the early industrializers did. Second, reductions in income equality will require profound structural changes. Liberal reform measures are constrained by the dependent production system and the internal class configurations. Therefore within the market rules of the game, income redistribution to labor would threaten to reduce surplus production allocated to export and investible surplus needed for capital accumulation.

A growth strategy committed to sharp reductions in inequality would involve structural changes in the purpose of production and the labor process along the following lines. First, production for use would replace the system of production for profit. A central plan with control over the general lines of development must be integrated with decentralized, largely self-reliant local units. Given the widely divergent latifundia/minifundia productivity rates, land reform will go some way to creating reorganizational space, increasing labor productivity and providing basic food needs. However, breaking down the urban/rural disparities also requires emphasis on rurally located light industries, produced in small to medium sized enterprises and applying traditional and intermediate technologies. The development of an integrated local and national social organization committed to producing a basic consumption bundle for the mass of the population can put people to work not in expanding exports of primary products or production of upperclass consumption goods, but in producing irrigation facilities, nutritious food, preventative health care, transportation systems, and ecologically sound and publicly accessible technologies.

Keynesian reforms fall short of production for use and tend to be self-defeating in the context of poor countries. For example, large scale public employment programs within a system dominated by production for profit can only channel resources into endeavors that do not threaten

entrenched private interests; as a consequence productivity rates are usually low which, in turn, soon generate inflationary and foreign exchange pressures and a call for government retrenchment. To the extent products are modified and production is reorganized to meet the needs of the underprivileged, both the scope of inequality and the long-run dependency upon foreign markets and multinational corporations will be reduced. Clearly the primary obstacle to production for use is the political power held by owning classes.

Second, the wage-labor system must be replaced with a cooperative labor system. The former is defended by neoclassical economics on the grounds that individual, external incentives, usually in the form of money, are necessary to motivate labor. Rooted in the individualistic theory of human nature, it presumes that an individual factor's contribution to a social process can be isolated and measured and ignores the inner connections between income distribution, incentive systems and worklife organization. An alternative *social* theory of human nature holds that substantial inequality does not reflect universal laws necessary to motivate savings and labor, but rather privileged access to property and the imperatives of labor motivation under alienating work regimes.

A social theory of human nature emphasizes that the dichotomy between decision makers and decision takers leaves the worker with virtually no control over what will be produced, how it will be produced, and little sense of contributing useful goods or services to the community. The separation of owners-managers from workers, of mental from manual, relegates the individual worker to mindless tasks, crushing the opportunity to develop her/his human capacities to reflect, design, invent, and to relate interpersonally in emotionally complex but self-conscious ways. In short, the inherent fragmentation, isolation, and instrumentality of work dehumanize social relationships. The motivation of labor under such arrangements does require individual and external incentives which stand outside the labor process.

Until income distribution analyses emphasize such ties between worklife arrangements and incentive systems, they will be specific in nature and fundamentally conservative in implication. The move toward equality will require a broader blend of external, individual, internal, and social incentives. Democratizing worklife, by breaking down the divisions between manual and mental, makes jobs more fulfilling as ends in themselves. In this way, worker participation in decision making reduces the need for external incentives. Similarly workers who have gained a stake in the production process and in the community have reason to shift from passive conformist attitudes to motivated agents sharing in the task of expanding production. Worker participation in the design of production, besides reducing the burden placed on external and individual incentives, enhances the inventive capacities of workers and thereby worker productivity.

Seen from this perspective, it is little wonder that systems which relegate workers to mindless tasks require coercion, such as the threat of impoverishment to extract labor effort. It also helps to explain the progress of China and Cuba. Both have broken down traditional market oriented divisions of labor and integrated workers into decision-making capacities. Both have motivated enormous labor effort by creating an institutional context of direct worker participation in plant organization and by fostering a social environment of interdependence and reciprocity as opposed to atomization and individualism.

Clearly the democratization of worklife

will also require production for use. Workers must have a say in determining what will be produced if social incentives are to be meaningful. Once more, moves toward workplace democratization will be strongly resisted by private owners for, to them, a worker capable of designing production alternatives is potentially capable not only of competition within the prevailing economic order but of designing an alternative economic order.

These reforms go beyond technical adjustments in governmental policies; they point to changes in the relationships of production and the ideology of individualism. The latter is itself a product of institutional arrangements that preclude collective social action. For the underclass it is an ideology of frustration and impotence which leads victims to blame themselves for social injustices and to perceive the prevailing order as immutable and the existing conditions as hopeless.

Obviously, alterations in the purpose and process of production and in the sustaining ideology will require a social movement to secure implementation and will be resisted with all the power at the disposal of the dominant class. But the political obstacles to breaking with the traditional division of labor are not in themselves an argument against our interpretation of inequality or our proposals for change as much as an argument against the system which obstructs such programs.

A final note. The deep-seated nature of social change required to move the world economy toward greater equality can elicit a sense of pessimism. It shouldn't. The task of science is to make transparent the critical forces that are not obvious, to expose the underlying structural relationships. Once the forces producing inequality, for example, are made evident it is clear that the barriers to creating a better world are not imposed by nature, but are social problems subject to social action. The potential offered by rational, informed, collective social action will not be realized without struggle. That struggle must be informed by a proper identification of the real barriers to a more equitable world order.

# INTERNATIONAL INTERDEPENDENCE: IMPLICATIONS FOR POLICY

# Trade and Employment Effects in the United States of Multilateral Tariff Reductions

By ROBERT E. BALDWIN*

This paper summarizes certain aspects of my research into the trade and employment effects in the United States of a significant multilateral reduction in trade-distorting measures by the world's major trading nations. The study differs from earlier investigations into this question such as those by Giorgio Basevi (1968), Stephen Magee (1973), Robert Stern (1964), and Beatrice Vaccara and Walter Salant (1960) in that the industry breakdown is much more detailed and the consequences of a multilateral tariff reduction on both *U.S.* export and import-competing industries is taken into account. By estimating not only the net trade and employment effects of a significant tariff reduction in over 350 industries but also in the 50 states and on some 14 occupational groups, it is hoped that the results will be useful for those who are now embarked on the so-called Tokyo round of trade negotiations within the framework of the General Agreement of Tariffs and Trade (*GATT*). Another novel feature of the study is the estimation of the net employment effects of multilateral tariff cuts under the assumption of flexible exchange rates.

## I. Analytical Framework

A key assumption of the analytical framework used in the study is that imports are imperfect substitutes for domestic production. The usual perfect substitution model may be adequate for dealing with agricultural products and raw materials, but it gives inconsistent results when empirical estimates of elasticities of import demand, domestic demand, and domestic supply are compared with the theoretical relationships among these variables that must hold in a perfect substitution model.[1] A more appropriate model for trade in manufactures is one in which imports and domestic production are imperfect substitutes and supply curves for each of these types of goods are infinitely elastic. The assumption of an infinite supply elasticity of imports is traditional in trade theory,[2] while the

* Professor of Economics, University of Wisconsin-Madison. Research for the paper was financed by the U.S. Department of Labor and the Ford Foundation.

[1] If domestic and import products are perfect substitutes then $e_m=(1+O_s/O_m)e_d+(O_s/O_m)e_s$, where $e_m$, $e_d$, $e_s$ are, respectively, the elasticities of import demand, domestic demand, and domestic supply and $O_s$ and $O_m$ are the levels of domestic output and imports, respectively. However, for most manufactured products, when empirical estimates of $e_m$ and $e_d$ are combined with actual $O_s/O_m$ ratios, the implied $e_s$ ratio associated with these figures is much smaller than the very high values of $e_s$ obtained from direct empirical estimation.

[2] Although an infinite import supply curve is generally assumed, recent evidence seems to indicate that it is upward sloping for the United States. For example, Peter Clark, "The Effects of Recent Exchange Rate Changes on *U.S.* Trade Balance," Treasury Conference, April 1974, found that a 1 percent decrease in the *U.S.* demand for finished manufactures caused foreign producers to reduce their prices by .32 percent. He also found that as foreign demand for *U.S.* exports increased, there was no change in the dollar price of *U.S.* exports.

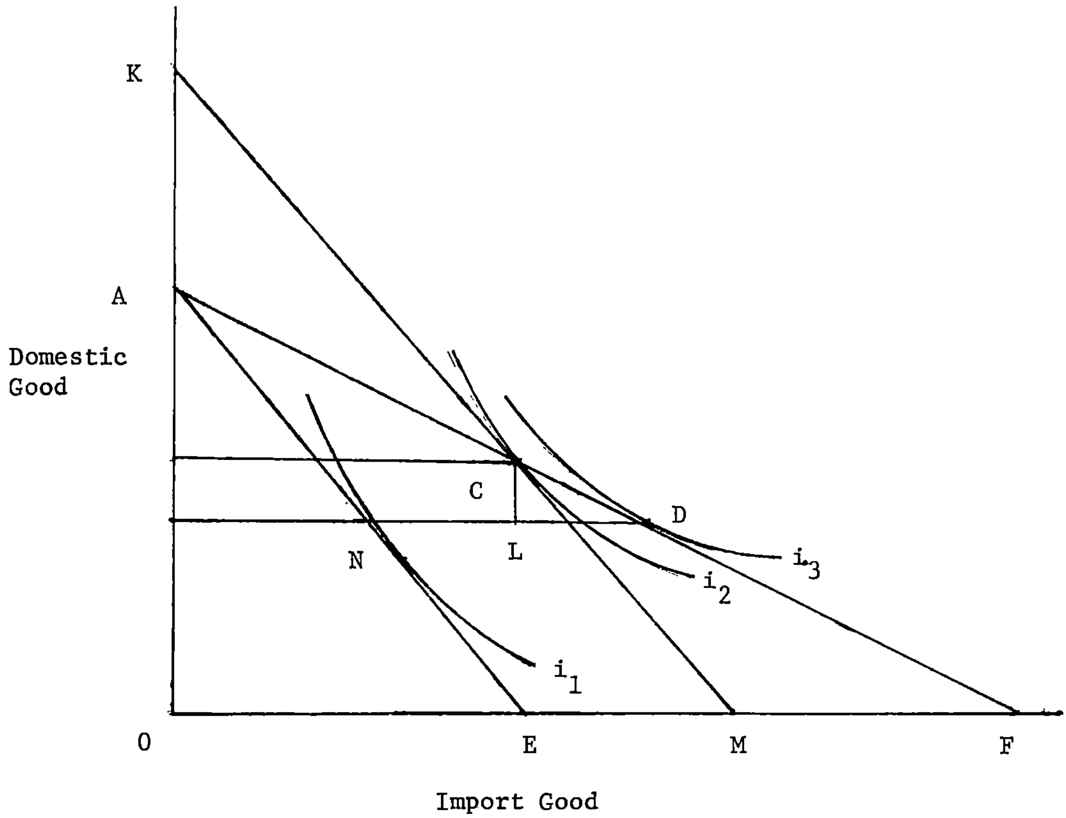



FIGURE 1

infinite supply elasticity for domestic production of manufactured goods is widely supported by empirical estimation, e.g., A. Walters (1963).[3]

Another simplifying assumption made in the analysis is that the compensated cross-price elasticity of demand is zero between any import good and all other goods except for the domestic substitute for this import good. Moreover, as is customary in tariff analysis in order to eliminate any macro effects, it is assumed that total government expenditures and tax revenues remain unchanged. Consequently, if government expenditures are unchanged, as will be assumed here, and import duties are changed, it is assumed that income taxes are changed in such a way as to hold total tax revenue unchanged.

These points can be better understood with the aid of Figure 1 in which a simple exchange situation is depicted. Let $OA$ be the country's supply of the domestic good, the slopes of $AE$ and $AF$ reflect the domestic price and the fixed international price, respectively, of the import good, and $i_1$, $i_2$ and $i_3$ be three community indifference curves. Postulating that consumers are identical in every respect, the point $N$ indicates the optimal combination of the domestic and import goods if a duty is imposed without any offsetting change in either taxes or government spending. However, if income taxes are also reduced at the same time by a sum equal to the increased tariff revenues measured in terms of the domestic good

[3] In later work a less than infinite elasticity of domestic supply will be used for agricultural and mineral products, but at this stage the infinite elasticity assumption is employed for these products, too.

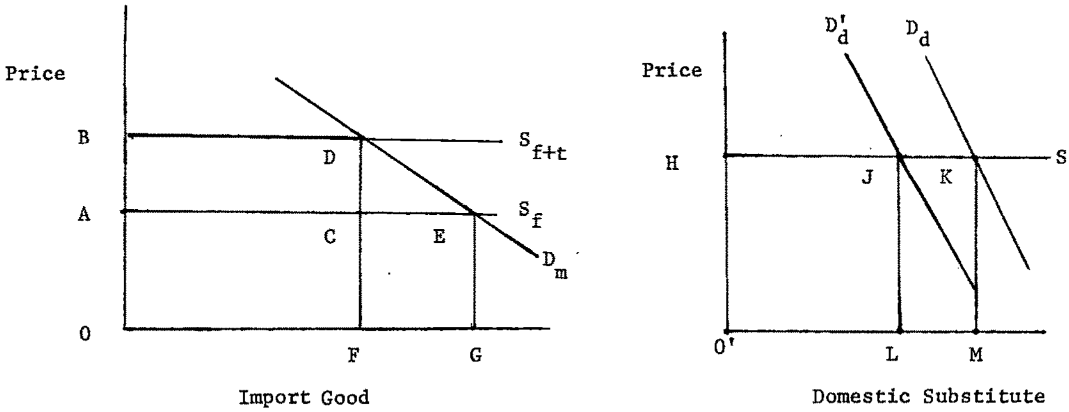



FIGURE 2

($AK$ equals both the tariff proceeds and the additional disposal income due to the cut in the income tax), consumers divide the additional income between the two goods in such a way that the point $C$ is reached. Consequently, to take the opposite case where the tariff on the import good is removed and the income tax is increased at the same time, consumers trade from the point $A$ along the line $AF$ until the point $D$ is reached.

Since the line $AF$ passes through $C$, consumers are able to purchase the same quantities of these two goods. In other words, apparent real income remains the same. Although consumers do not wish to substitute the import good for third goods, they do wish to purchase more of the import good and less of its domestic substitute. This substitution effect from $C$ to $D$ in Figure 1 indicates the increase in actual real income in the move to free trade conditions.

A representation of these relationships using ordinary demand and supply curves is shown in Figure 2.[4] In the figure, $D_m$ and $D_d$ are income-compensated demand curves; $S_f$ and $S_{ft}$ are the foreign supply of imports without and with, respectively, the import tax; and $S_d$ is the domestic supply curve of the import substitute. The area $ABDC$ is equal to the tariff proceeds as well as the decrease in the income tax and corresponds to $AK$ (or $EM$) in Figure 1. When the duty is eliminated and the sum paid to foreign producers increases by $FCEG$ (equals $LC$ in Figure 1), domestic expenditures on the local substitute decline by an equal sum, namely, $LJKM$ in Figure 2 or $LC$ in Figure 1.

## II. Data Sources

The trade and tariff data needed for the study are available for 1970 and 1971 from the secretariat of $GATT$. This organization obtained tariff and trade information on a tariff-line basis from its major trading members and then classified it according to the Brussels Trade Nomenclature ($BTN$) at a 4-digit level.[5] The industry breakdown employed corresponds to the 367-sector delineation of the 1967 sector input-output table prepared by the U.S. Department of Commerce. Intermediate-use coefficients from this study were also employed.

Five additional sets of data are utilized in estimating the trade-balance impact of multilateral duty reductions and then the

[4] One difference is that domestic production of the domestic good is permitted to vary.

[5] The countries covered are: the United States, Japan, Canada, Austria, New Zealand, Finland, Switzerland, Sweden, Australia, Norway and the nine European Community ($EC$) members, namely, France, Germany, Italy, Belgium, Luxembourg, Denmark, the United Kingdom, and Ireland.

employment changes associated with these trade effects. They are: import and export demand and supply elasticities for the 310 trading industries in the 367 input-output table; employment coefficients classified on the same industry basis; a breakdown of employment in each industry by skill groups, a breakdown of employment in each industry by state; and finally a set of price deflators to put the 1971 trade figures on a comparable basis with the 1967 input-output and labor-output coefficients used in the study.

Trade and employment effects were calculated using five alternative sets of import and export of demand elasticities.[6] The preferred set (Set I) is based on the import and export demand elasticities calculated by Margaret Buckler and Clopper Almon (1972) for some 75 manufactured goods and import demand elasticities calculated by Stephen Magee (1970) for crude foodstuffs, crude materials and manufactured food. Another set (Set II) consists of the Buckler-Almon elasticities modified by estimates for 20 industries calculated by the staff of the International Trade Commission. The remaining three elasticity sets consist of a fourfold grouping of one-digit Standard International Trade Classes and are based on a survey of the literature on trade elasticities by Robert Stern (1975). One set is termed the "best" by Stern; another set (Set IV) is made up of import demand estimates of previous investigators that are on the low side and export demand elasticities on the high side; and a final set (Set V) consists of high import and low export elasticities.

[6] The desired but unavailable (in the required detail) elasticity data are import demand elasticities for each participating country. Rather than use *U.S.* import demand elasticities for all countries, *U.S.* export demand elasticities are utilized, since the aggregative results are substantially the same and the different competitive position of an industry on the import versus the export side is taken into account.

Labor coefficients for the 367 sectors of the 1967 input-output table were kindly supplied by Clark Bullard of the University of Illinois.[7] A breakdown of the labor coefficients into 14 different skill groupings was supplied by the *BLS,* while sector employment by state was obtained from the 1970 Sample Census of Population.[8] Finally, in order to express 1971 trade figures in terms of the 1967 prices used for the input-output and labor coefficients, exports and imports classified into a 4-digit level of the Standard Industrial Classification (*SIC*) were deflated by 4-digit unit values obtained from the Bureau of Economic Analysis of the U.S. Department of Commerce. After this deflation the trade figures were then classified on the basis of the 367 sectors of the 1967 input-output table.

### III. Estimates of the Trade and Employment Effects of Tariff Reductions

The aggregate trade and employment effects in the United States of a 50 percent, multilateral tariff reduction are presented in Table 1. The Buckler-Almon-Magee elasticities are utilized in calculating these effects. In addition, certain products have been excluded from the tariff-cutting process, namely, *U.S.* exports to the European community (*EC*) of agricultural products subject to quantitative restrictions in this country, all *U.S.* exports and imports of textiles covered by the international textile agreement, and *U.S.* imports of oil. All of these commodities are subject to nontariff barriers

[7] Some modifications in these coefficients were made after consulting with the Bureau of Labor Statistics (*BLS*) in the U.S. Department of Labor. The figures also are average rather than marginal labor coefficients. However, if short-run supply curves are in fact horizontal, there will be no difference in these two types of coefficients.

[8] Employment by state is divided into only 19 industries.

TABLE 1—EFFECTS ON U.S. TRADE AND EMPLOYMENT ON A 50 PERCENT LINEAR TARIFF REDUCTION

| | All Industries | Manufacturing |
|---|---|---|
| Change in exports (millions of $) | 1,750 | 1,591 |
| Change in imports (millions of $) | 1,746 | 1,717 |
| Net trade change (millions of $) | 4 | −126 |
| Export-related employment change (man-years) | +136,000 | 116,400 |
| Import-related employment change (man-years) | −151,200 | −148,100 |
| Net change in employment (man-years) | − 15,200 | − 31,700 |

that operate independently of tariffs. Cutting duties on these items will not increase trade.

While a 50 percent duty reduction leads to changes in exports and imports that each exceed $1.5 billion, the net trade change is a negligible $+4 million, for all industries and $–126 million for manufacturing alone. The net employment impact is also small: –15,200 man-years for all industries and –31,700 man-years for manufacturing.

As previously noted, estimates of trade and employment effects from a 50 percent tariff cut were made with four other elasticity sets. Table 2 gives these results. Unless one combines import demand elasticities taken from the high side of the range of such figures that various investigators have found together with export demand elasticities taken from the low side of the range of estimates for these elasticities, i.e., Set V, the figures in Table 2 show that aggregative trade and employment effects of a 50 percent reduction of tariffs are small or moderately favorable. The remainder of the results reported in this paper are based on the Buckler-Almon-Magee elasticity estimates, i.e., Set I.

Some indication of the distribution effects of tariff reductions can be obtained by breaking down aggregate employment changes into various skill groups. Dividing net employment changes into fourteen occupational classes yields the following percentage changes in labor requirements: research and development workers at + .14; professional and technical workers (production related) +.08; other professional and technical workers 0; management and administrative employees (production related) –.03; other management and administrative employees 0; craftsmen –.05; sales workers –.03; clerical workers –.02; operatives –.14; laborers (except farm) –.08; service workers –.01; private household workers 0; and farm laborers

TABLE 2—AGGREGATE U.S. TRADE AND EMPLOYMENT EFFECTS OF A 50 PERCENT TARIFF CUT UNDER VARIOUS TRADE ELASTICITIES

| | Set I | Set II | Set III | Set IV | Set V |
|---|---|---|---|---|---|
| Net trade effect | | | | | |
| (in millions of $) | | | | | |
| All industries | 4 | 161 | −310 | 1,493 | −2,932 |
| Manufacturing | −126 | 31 | −343 | 1,328 | −2,834 |
| Net employment effect | | | | | |
| (in man years) | | | | | |
| All industries | −15,200 | 900 | −37,300 | 113,300 | −266,400 |
| Manufacturing | −31,700 | −15,600 | −40,900 | 93,100 | −251,200 |

and foremen +.45. These figures confirm once again that the basis of the *U.S.* comparative advantage rests on the existence of a relatively abundant supply of research and development workers (who give the United States a technological jump on other countries), other professional workers, and agricultural resources. The demands for other skill groups decline somewhat with the reduction being the largest for semi-skilled (operatives) and nonfarm laborers.

While overall trade and employment effects of a 50 percent tariff cut are small under the most likely sets of trade elasticities, this does not mean that no industry is significantly harmed or benefited from the duty reductions. If one-half of one percent of an industry's labor force is arbitrarily chosen to divide industries that are or are not significantly effected, the list of industries in which there are employment changes of one-half of one percent or more consists of 54 import-sensitive industries and 19 export-oriented industries. The employment change for the import sensitive industries is 48,000 and for the export-oriented industries, 34,000.

In considering whether to phase cuts in import-sensitive industries over the full ten years permitted by the Trade Act of 1974, an initial consideration might be whether the annual growth rate for labor in an industry more than offsets the labor force decline associated with a ten-year phasing arrangement. If it did, there would be no reason for not reducing duties in the industry by the full 50 percent. If average annual growth rates for labor that prevailed from 1958–71 are representative of the next ten years, the growth factor eliminates unemployment problems caused by the tariff reductions in 34 of the 54 import sensitive industries. The remaining 20 consist of such industries as lace goods, tire cord fabric, cordage and twine, industrial leather, nonrubber footware, glass products, pottery products, lead and zinc, textile machinery, electronic tubes, and artificial flowers.

Industries that benefit in employment terms by one-half or more percent include: paperboard mills, computing machines, electrical measuring instruments, semiconductors, aircrafts and aircraft equipment, and scientific instruments.

The regional distribution is also of interest to trade negotiators, and therefore a state by state employment impact of a 50 percent duty reduction has been estimated. The 50 percent multilateral tariff reduction brings about an employment decline in 34 of the 50 states. However, these losses are very small. For example, the states in which labor requirements fall by more than 1000 man-years are Illinois, Massachusetts, Michigan, New York, and Pennsylvania. In relative terms, employment losses are the highest in four New England states—Massachusetts, Maine, New Hampshire, and Rhode Island, whereas the relative employment gains are the greatest in Kansas, Minnesota, North Dakota, and South Dakota.

All of the previously reported results are based upon a fixed exchange-rate model. However, it is possible to estimate the exchange-rate change required on the part of the United States to eliminate any deficit or surplus pressures on the trade balance as a result of tariff cuts. For elasticity Sets I-III, the required percentage exchange-rate changes needed for a 50 percent tariff cut on all industries are within a range of +.003 and –.688. The extreme elasticity assumptions used in Sets IV and V yield +2.43 and –2.46, respectively. There are still, of course, labor demand effects even though there is no trade-balance impact. These changes are, however, small for all elasticity sets.

## IV. Conclusions

The main conclusion emerging from the study is that the United States can participate in a substantial tariff-cutting negotiation without causing significant adverse trade and employment effects in the country. Even without taking into account exchange-rate changes, any adverse trade and employment effects are small except under one extreme set of trade elasticities. When exchange rate variations are taken into account, net aggregate trade changes are eliminated and total employment shifts become minimal even under the most adverse elasticity assumptions.

Not only are aggregate economic effects of a significant tariff-cutting exercise small, but the effects on individual industries, on various occupational groups, and on employment in different states are minimal in most cases. For example, as previously noted, normal industry growth can handle any adverse employment impact, in all but 20 industries, if the reductions are staged over a ten-year period. Employment changes both by skill group and by state are insignificant, especially if the cuts are staged over a ten-year period.

## REFERENCES


**G. Basevi**, "The Restrictive Effect of the U.S. Tariff and Its Welfare Value," *Amer. Econ. Rev.*, Sept. 1968, *58*, 840–52.

**M. Buckler and C. Almon**, "Imports and Exports in an Input-Output Model," *Amer. Statist. Assn. Proc.* 1972.

**S. P. Magee**, "Prices, Incomes, and Foreign Trade," Conference on Research in International Trade and Finance, Princeton Univ., Working pap. no. 19, Mar. 30–31, 1973.

———, "A Theoretical and Empirical Examination of Supply and Demand Relationships in U.S. International Trade," Study for the Council of Economic Advisers, Oct. 1970.

**R. M. Stern**, "Price Elasticities in International Trade: A Compilation and Annotated Bibliography of Recent Research," prepared for the Office of the Special Trade Representative, Washington, mimeo, March 1975.

———, "The U.S. Tariff and the Efficiency of the U.S. Economy," *Amer. Econ. Rev.*, June 1964, *54*, 459–70.

**B. Vaccara and W. Salant**, *Import Liberalization and Employment*, Washington 1960.

**A. A. Walters**, "Production and Cost Functions: An Econometric Survey," *Economica*, Jan.-Apr. 1963, *31*, 1–66.

# American Multinationals and the U.S. Economy

By THOMAS HORST*

The impact of American multinational firms on the United States economy and its foreign relations is the subject of a study, *American Multinationals and American Interests*, which Fred Bergsten, Theodore Moran and I have recently completed for the Brookings Institution. If condensing two years of work by three authors into one book was difficult, summarizing my contribution here is all but impossible. But assuming that the effort would generate more interest than a careful dissection of the foreign tax credit, let me describe our more important findings in the areas of international trade, income taxes, and antitrust.

## I. Trade Issues

The impact of American investments abroad on *U.S.* exports and imports is certainly a controversial issue. Whatever interest it may hold in its own right is overshadowed by its implications for the export of American jobs, the distribution of income between American labor and capital, and the balance of payments. In approaching this complex issue, we focused on the microeconomics of foreign trade and investment rather than on the general-equilibrium linkages. An American multinational chooses—explicitly or implicitly—whether to supply its foreign and domestic markets with goods manufactured at home or overseas. When a company like International Business Machines (*IBM*) decides to manufacture its computers for its European customers in Europe, it displaces *its own* exports of computers from the United States. While data on the exports, imports and foreign affiliate sales of individual American investors are exceedingly hard to come by, industry statistics are comparatively easy to find and seemed altogether appropriate for demonstrating the substitution of subsidiary production for American exports.

While this notion of direct substitution has obvious intuitive appeal and is implicit in the labor unions' charges that American investors export their jobs, we had no success whatsoever in demonstrating statistically that it was a major factor in shaping the flows of *U.S.* exports and imports. If one industry's exports and foreign subsidiary sales are compared to another's, one often finds that the industry which exports more also produces more overseas. This simple correlation does not imply, however, that exporting stimulates foreign production or vice versa. Often one industry is more technically sophisticated than the other, and its technical capability assists its exports *and* its foreign production. What surprised us was that when we used common statistical procedures to correct for differences in technical sophistication and other spurious factors, we still could not demonstrate any clear or obvious tradeoff between exporting and foreign subsidiary production. There was no evidence that an increase in foreign subsidiary production was normally accompanied by a reduction in *U.S.* exports or, for that matter, an increase in *U.S.* imports. In fact, there was some evidence that a *modest* amount

* The Fletcher School of Law and Diplomacy, Tufts University.

of foreign investment might be prerequisite to exporting. But ignoring that and focusing on the industries with substantial exports or overseas markets, we found no statistically significant relationship across industries between foreign investing and *U.S.* exports or imports.

What conclusion should we draw from the absence of any such relationship? That our exports are never displaced or our imports never enhanced by foreign production? That tariffs, transport costs, international differences in labor costs and exchange rates have no impact on the location of a multinational's production? Most certainly not. But our negative findings do suggest that the implications of subsidiary production for *U.S.* exports or imports have been overstated and/or that the ways in which foreign affiliate production may complement *U.S.* exports have been underestimated. Although economists often ignore the sort of market imperfections on which the complementarity may be based, the histories of multinational corporations or industries are more revealing. When American firms invest abroad, they not only produce goods which hypothetically might be produced in the United States, but they also undertake a wide variety of nonmanufacturing activities to expand the foreign market for their products. Direct selling, advertising, wholesale and even retail distribution, adapting American products to local market conditions, follow-up maintenance and repair all play an indispensible role in broadening the market for the firm's products. In labeling these activities "ancillary" when they are performed by an American manufacturer, it has been all too easy to ignore their significance.

Our conclusions are simple. Published foreign investment statistics reflect a heterogeneous combination of manufacturing and nonmanufacturing activities, and while the foreign *manufacturing* may be an alternative to American production, the *nonmanufacturing* activities tend to promote *U.S.* exports by expanding the market for *U.S.* goods. We would not risk belaboring this point were it not for the regretable tendency of such distinctions to be the first casualties of public policy debates over foreign investment. The Burke-Hartke Bill, which was supported by the *AFL-CIO* and finally rejected in 1973 by the Congress after a long and bitter debate, proposed to repeal the foreign tax credit for *all* foreign investment. By subjecting foreign investment income to double taxation, the labor unions hoped to slow down or reverse the export of jobs. We conclude that such a policy in failing to discriminate between one type of foreign investment and another could easily do more harm than good to the U.S. trade balance. The depreciation of the U.S. dollar did far more to reverse the export of jobs than the Burke-Hartke Bill ever could have.

## II. Tax Issues

Under current United States tax policy, foreign investment income is subject to the corporate income tax, but not until that income is formally paid as a dividend to the American investor, and then a tax credit is given for foreign income taxes. Even if this policy has had an ambiguous impact on *U.S.* exports and imports, it has had two unfortunate side effects. It has allowed foreign governments to collect the lion's share of taxes on foreign investment income, a phenomenon which is highly significant in any reckoning of the national gains from international investment. Secondly, it has encouraged American investors to undercharge their foreign affiliates for intra-company exports and for mutually beneficial research and development programs and other joint expenses. In the end the United States has

TABLE 1—ESTIMATED IMPACT OF ELIMINATING DEFERRAL FOR MANUFACTURING FIRMS, 1974
(Millions of dollars)

| Variable Affected | Impact Assuming No Change in Intra-Firm Financial Practices | Estimated 1974 Value | Percentage Change |
|---|---|---|---|
| Total Domestic Assets | 2,436 | 338,400 | .7% |
| Total Foreign Assets | −3,216 | 140,400 | −2.2% |
| Intra-firm Financial Transfer | −11,046 | 28,080 | −39.3% |
| Consolidated After-Tax Earnings | −755 | 43,123 | −1.8% |
| *U.S.* Income Taxes[a] | 1,233 | 28,815 | 4.2% |
| Foreign Income & Withholding Taxes | −433 | 9,569 | −4.5% |

[a] *U.S.* income taxes include taxes paid on parents' domestic income.

little or no share in the taxes paid on the income generated by American investments abroad and may even lose some tax revenues on domestic income to overseas governments.

As an aid in evaluating various tax reforms, we have constructed a microeconomic model of a multinational firm and used various estimates of the model's parameters to simulate the multinational's response to possible changes in United States tax policy. In Table 1 we show our estimates of the impact of eliminating deferral for manufacturing investors in 1974. Without the deferral of foreign investment income, we estimate that the total assets of foreign manufacturing affiliates would be reduced by $3.2 billion, which is 2.3 percent of their estimated total value. The total domestic assets of the multinationals might rise by $2.4 billion, which is only 0.7 percent of these firms' domestic assets. That domestic investment rises by less than foreign investment falls reflects the loss of investable funds through higher tax payments to the U.S. Treasury. The primary reason why eliminating deferral has so small an impact on foreign and domestic investment is that foreign governments collect income and withholding taxes amounting to 45 percent of foreign investment income. If deferral were repealed, the total taxes imposed on foreign investment income would not exceed 48 percent, the statutory *U.S.* tax rate. If repealing deferral adds only 3 percent to the total taxes on foreign investment income, the impact on the location of investment is apt to be small.

But this is not to say that the United States has little to gain by eliminating deferral. Far from it. We estimate that the total taxes paid by American investors to the U.S. Treasury (including taxes on domestic income) would rise by $1.2 billion, a 4.2 percent increase in those taxes. Eliminating deferral has a comparatively larger impact on U.S. taxable income and tax payments than on domestic investment (a 4.2 percent gain versus a 0.7 percent gain). This striking difference reflects our presumption that the elimination of deferral would encourage American investors to finance more of their overseas investment with locally borrowed funds than they now do. If the subsidiaries borrow in local capital markets, the interest is deductible from their taxable income and not from their American parents' domestic income. The gains from encouraging American investors to refinance their foreign operations constitute a major incentive to repeal deferral.

The tax gains of this particular reform do not end here. Deferral encourages firms to undercharge their foreign affiliates for research and development programs undertaken in the United States. Although American investors often spend as much as 5 percent of their domestic sales receipts on research and development programs, only a small portion of the total cost is charged back to the foreign affiliates. In 1973, for example, total royalties, management fees, rentals and all other such charges amounted to only 1.1 percent of manufacturing affiliates' sales. If repealing deferral encouraged firms to raise these charges by another 1.1 percent, the tax gain for the United States in 1974 would have been almost $1 billion—a gain over and above that shown in Table 1.

So while we favor the repeal of deferral, we believe that the primary benefits are quite different from those stressed by earlier writers. The principal gain derives from the tax benefits of encouraging American investors to alter the capital structure of their overseas affiliates and to increase the portion of taxable income allocated to the *U.S.* parent. The more familiar gains—the more efficient pattern of international investment and the additional *U.S.* taxes collected on subsidiaries retained earnings—may pale by comparison.

### III. Antitrust Issues

As Stephen Hymer, Raymond Vernon, Richard Caves, and several other authors have noted, American multinationals are anything but the atomistic competitors envisaged by classical economic analysts. Multinational firms tend, if anything, to be the largest firms in their industries, and those industries are distinguished by high advertising, high research and development spending, substantial economies of scale and other wellsprings of monopoly power. The primary connection between multinationalization and monopolization is simple: the technological and marketing advantages which allow a few firms to dominate the market also serve as a springboard for entering foreign markets.

Our interest was not in reconfirming this primary linkage between foreign investment and market power, but in determining whether foreign investment reinforces or entrenches the multinationals' market position in the United States. Our suspicions were aroused not by the opportunities a multinational may have to use its foreign earnings to finance predatory attacks on its domestic competitors (as Richard J. Barnet and Ronald E. Müller argued in *Global Reach*), but by the possibility that the underlying sources of market power—patents, brand differentiation, vertical integration—would be augmented by foreign investment.

A prima facie case for this hypothesis is easy enough to make—the multinational companies themselves often claim that only through their large, global operations can they afford the substantial research and development programs they undertake. The fruits of their research efforts help them fend off foreign competition when they enter foreign markets while enhancing their premier positions in the United States.

Whether the contribution of foreign investment to domestic market power is large or small is a much harder question to answer. We have examined Internal Revenue Service statistics on the domestic rates of return for firms of different sizes and in different industries. Industrial organization analysts reason that firms will be able to earn consistently higher profits than average only if they are protected in one way or another from new competition. A summary of our findings appears in Table 2 where we have cross-tabulated firms' domestic returns on capital according to whether they were

"multinational" or "domestic," whether they spent heavily on advertising or not, whether they employed many scientists and engineers or not, and whether they were large or small. Comparisons are made for 1966, a good year, and 1970, a bad one. By reading up and down the columns, one can see that firms spending heavily on advertising consistently showed higher rates of return than firms which advertise less. The advantages of technological effort are less striking, but apparent nonetheless. Large firms do not, however, consistently earn more than small firms do.

TABLE 2—CROSS TABULATIONS OF DOMESTIC RETURNS AS A PERCENT OF PARENTS' TOTAL ASSETS, 1966 AND 1970

| | Multinational Groups | Domestic Groups |
|---|---|---|
| *1966* | | |
| High Advertising | 17.0% | 13.8% |
| Low Advertising | 13.0% | 11.4% |
| *1970* | | |
| High Advertising | 14.3% | 10.8% |
| Low Advertising | 8.1% | 6.6% |
| *1966* | | |
| High Technology | 16.7% | 13.2% |
| Low Technology | 12.6% | 11.5% |
| *1970* | | |
| High Technology | 11.0% | 5.7% |
| Low Technology | 8.7% | 7.1% |
| *1966* | | |
| Large Firms | 14.9% | 11.6% |
| Small Firms | 13.8% | 11.7% |
| *1970* | | |
| Large Firms | 9.7% | 6.2% |
| Small Firms | 9.6% | 7.3% |

*Note:* Domestic Net Returns equal net income before taxes plus interest paid on debt less foreign dividends and tax credits. Multinationals were those industry-size-class groups whose foreign dividends and tax credits exceeded 1 percent of parents' assets; high advertisers were those groups spending amounts more than 5 percent of parents' assets on advertising; high technology *industries* were those whose 1970 employment of scientists and engineers exceeded 5 percent of their total employment; large firms had more than $250 million in total assets. The "groups" of American corporations consisted of all firms of a given size class (e.g., $50-$100 million in total assets) in one of seventy-five "minor" industries (e.g., meatpacking). The source of these statistics is the Internal Revenue Service, *Sourcebook of Statistics of Income, Corporations, 1966* and *1970*, obtained from the National Archives in machine-readable form.

The more interesting relationship, so far as we are concerned is apparent in reading across the rows. Now we are comparing firms spending comparable amounts on advertising, employing comparable numbers of scientists and engineers, or of comparable size to determine whether those with foreign operations earn higher returns on their *domestic* investments than those without overseas investments. What is striking about these horizontal comparisons is that for high advertisers and low, large and small employers of scientists and engineers, and large firms and small, *the multinational groups earned significantly higher domestic returns than domestic groups in 1966 and 1970.* (The relationship also held for each of the seven years 1965–71.) The possibility that foreign investment significantly increases the domestic monopoly power of multinational firms clearly should not be dismissed out of hand.

Unfortunately, this sort of monopoly power hits American antitrust policy right where it is weakest. Antitrust policy has had its greatest successes in undoing price-fixing and market-sharing arrangements among would-be cartels and in preventing horizontal and vertical mergers among large firms. Its greatest failures have come in attacking market power inherent in economies of scale, technological knowhow and other basic elements of industry structure. These failures may be due in part to a lack of legal resources or political support for antitrust prosecution of large firms, but are also a reflection of a troublesome tradeoff between competition and economic efficiency. Even if the Justice Department or the Federal Trade Commission can prove that a technologically advanced firm has illegally monopolized a market, judges are often reluctant to split the company into several

smaller, independent competitors for fear of sacrificing substantial economies of scale the firm achieves in production or in developing new technology. Foreign investment tends to reinforce domestic market power in those very cases where antitrust authorities are least willing and able to act.

Every cloud has its silver lining, and this one is no exception. In trying to find an effective remedy for this type of monopoly power, antitrust authorities often look for natural ways to split dominant firms into self-sufficient entities, just as a jeweler looks for a natural way to split a diamond. In this search we would hope that antitrust authorities would not overlook the obvious opportunities to split the large and more sufficient foreign subsidiaries from their American parents. Although these subsidiaries are not current participants in *U.S.* product markets, they are often credible entrants. Is *IBM* Europe not a credible contender in the *U.S.* computer industry? Is General Motors, Canada not a potential entrant in the United States auto industry? In fairness to the American parent, we presume that it would be allowed to compete with its erstwhile affiliates in overseas markets, just as they would compete with their erstwhile parents in the domestic *U.S.* markets. In ordering the cross-licensing of critical patents and otherwise fashioning an effective remedy, the objective should always be to promote competition, not to emasculate one national firm or the other. Multinationals may thus derive competitive advantages over their domestic competition from their foreign operations, but those foreign operations may be a solution to, and not just a cause of, the problem of market power.

REFERENCES

R. J. Barnet and R. E. Müller, *Global Reach: The Power of Multinational Corporations,* New York 1974.

C. F. Bergsten, T. Horst and T. Moran, *American Multinationals and American Interests,* Washington, forthcoming.

# Increasing International Economic Interdependence: the Implications for Research

*By* C. Fred Bergsten and William R. Cline*

Increasing economic interdependence can perhaps best be defined as a growing impact of external events on national economies. One of its major implications for research is that all national economic models must be opened to encompass meaningfully the foreign sector.

Economists in most countries have opened their models long ago, but most American economists have not. Yet the share of exports in the *U.S.* economy has doubled in the last fifteen years, and the share of imports has doubled in just seven years. These *U.S.* ratios (about 7 percent, and still rising rapidly) are now only slightly below the same ratios in Japan and the European Common Market as a group (about 9 percent, and fairly stable).

In addition, about one-third of the profits of *U.S.* corporations now derive from overseas activities, primarily their foreign direct investments. If these profits are taken into account along with trade, the *U.S.* economy has probably become more open than Japan or Western Europe (as a unit) in quantitative terms. Even the absolute numbers are impressive: the *U.S.* trade balance, excluding the effects of the rise in oil prices, strengthened by at least $35 billion from 1972 through 1975—without which our gross national product would have been over two percentage points lower.

This openness has become critically important for *U.S.* economic policy. The overvaluation of the dollar was adding almost one full percentage point to the *U.S.* unemployment rate in 1971. External factors accounted for one-quarter to one-half of the rise in the rate of inflation from 1972 through 1974 (Richard Berner, Peter Clark, J. Enzler and Barbara Lowrey). The United States can no longer export its internal inflation through the operation of a dollar-based system of fixed exchange rates as it did to a significant degree in the late 1960's (Bergsten 1976).

Oil and food are only the most obvious examples of external shocks to particular *U.S.* economic sectors. About 20 percent of *U.S.* industrial production and over one-third of *U.S.* farm output are exported. The United States already imports more than 50 percent of its needs of nine of the thirteen key industrial raw materials. Closed models can no longer accurately depict or forecast the *U.S.* economy at either the macroeconomic or sectoral levels.

## I. The Effects of Foreign Economic Policy Instruments

This increasing openness of national economies has increased the importance of the instruments of foreign economic policy, as countries seek both to fend off negative effects of outside events and to harness such events to their positive objectives. New research is needed on a variety of aspects of these instruments.

Countries have traditionally sought to export unemployment to their neighbors through exchange-rate depreciations, ex-

* The Brookings Institution.

port subsidies and import controls. But in 1973–74, as inflation superseded unemployment as the primary internal problem in most countries, there were widespread efforts to export inflation through opposite types of external economic policy—exchange-rate *appreciations, export* controls, and unilateral import *liberalization*. For example, several countries (including Austria, Germany and Norway) revalued their exchange rates though they were not running payments surpluses and Japan spent at least $5 billion to keep the yen from depreciating, just a few years after many of the same countries had bitterly resisted massive international pressure for similar steps when they were running huge surpluses. Numerous countries (including some of the most protectionist such as Australia, Canada and Japan) cut their tariffs across the board, and many others (including the United States on such politically sensitive items as sugar and meat) eliminated import quotas. Both industrialized and developing countries invoked export controls across a wide range of primary and manufactured products (Bergsten 1974).

If inflation once more becomes the chief internal concern of most countries, these efforts to use external policy to export it may again become major sources of international conflict. Alternatively, in a world of continued stagflation, we could expect national efforts to export simultaneously both inflation and unemployment (in different sectors, as with the recent *U.S.* controls on food exports and steel imports).

These developments raise major issues for economic analysis. One is the need for empirical evidence on the external corollary of the Phillips curve: How much inflation (unemployment) must be accepted as the price for reducing unemployment (inflation) through a devaluation (appreciation) of the exchange rate or imposition (liberalization) of import controls? Protectionist pressures have been restrained throughout the recent recession largely by fears of their simultaneous inflationary consequences, but economists have done little to define the issue rigorously and hence lay a foundation for sensible policy choices.

Tests are also needed of the effectiveness of these policy instruments in meeting their proclaimed objectives. For example, export controls reduce internal inflation by increasing domestic supply of the controlled product. But they may add to inflation in the short run by weakening the exchange rate. They may also be inflationary in the long run by dampening the elasticity of domestic supply; on the other hand, they may trigger increased supply in other areas as consuming countries seek to diversify their sources. They may encourage the adoption of export controls by other countries. Policy makers have little intellectual basis for addressing such issues, which are now arising with great frequency due to the acceleration of internal economic interdependence.

## II. Flexible Exchange Rates

International monetary arrangements stand at the center of the increasingly interdependent world economy, and have undergone dramatic structural change in recent years. The present international monetary system is extremely heterogeneous. Most countries retain fixed exchange rates. But the minority of countries whose exchange rates are floating account for over seventy percent of world trade. Data will be emerging over the next few years which will make possible analyses of several key issues regarding the effects and desirability of alternative monetary regimes, and a comparison with earlier periods in which fixity or flexibility dominated.

A major issue is the effectiveness of more flexible exchange rates in achieving balance-of-payments adjustment. The effects seem clearly positive for the United States and Japan, but less so for Germany and Britain. Country studies rank high on the research agenda.

So do studies of the systemic impact of more flexible rates. Some observers claim that exchange-rate changes can only raise world prices and do nothing to eliminate payments disequilibria. Others view flexibility as a major deterrent to world inflation, either because it "bottles up" inflation in individual countries or because depreciations trigger tougher stabilization policies than do losses of reserves—the analogue under fixed exchange rates. Answers to this issue must also encompass the impact, if any, of greater flexibility on levels of world trade and investment and on the demand for international liquidity.

Studies are also needed of the microeconomic effects of more flexible exchange rates. How do they affect foreign direct investment and multinational enterprises; is it a coincidence that Europe feared *Le Défi Americain* only when the dollar became seriously overvalued, and that ratios of foreign direct investment to trade are extremely high for countries whose exchange rates have historically been overvalued (United States, Britain) and low for countries whose exchange rates have been undervalued (Germany, Japan) (Bergsten, Thomas Horst and Theodore Moran)? Conversely, how do multinational enterprises affect the functioning of flexible exchange rates; Raymond Vernon agrees with Richard J. Barnet and Ronald E. Müller that they may render them largely impotent. Another microeconomic question is what exchange rate(s) should be used to test the existence of dumping, or government export subsidies which trigger countervailing duties; this issue is of enormous practical significance since billions of dollars of *U.S.* imports are currently under investigation on one or another of these counts.

A final question on floating rates is whether they have provided more or less stability than fixed rates. Using quarterly spot rates to compare the fixed rate period 1959–1—1973–1 with the floating rate period 1973–2—1975–1, Cline (1976) has found a smaller coefficient of variation during the floating rate period for the dollar rate of the deutschemark, the pound, and the yen, although not for the French franc. Experience is too limited to provide a definitive comparison, but the evidence to date suggesting greater stability under a flexible regime is impressive in view of the special pressures on the international monetary system during this period and in light of the common impression that floating rates have gyrated wildly. Further empirical tests should include examination of the appropriate specifications including periodicity, the measure of fluctuation, and the use of individual cross rates or variously weighted "effective" exchange rates.

### III. Welfare Effects of Trade Liberalization

One key measure of increasing international economic interdependence is the widespread increase in trade-to-*GNP* ratios (P. J. Katzenstein). Yet theoretical, and especially empirical, understanding of international trade flows remains inadequate for answering several central policy questions. These issues have become particularly important in light of two contrary developments: the resurrection of protectionist pressures in a number of countries, and the commencement of the "Tokyo Round" of multilateral trade negotiations (*MTN*) whose aim is the broadest and deepest trade liberalization in history. Do increased trade flows pro-

vide significant welfare gains? Would new trade barriers be very costly?

The traditional measurement of welfare gains from tariff liberalization considers the resulting gain in consumer surplus, net of transfers away from the producers' surplus and government revenue (but adding new government revenue on extra imports if some tariff remains). This measure inevitably yields small magnitudes; it is a fraction of the increase in imports, which itself is a fraction of the base level of trade, which in turn is a fraction of *GNP*. While the removal of very high tariffs generates substantial benefits because the welfare effect rises with the square of the tariff, the tariff structure in industrial countries today is generally low. Indeed, preliminary estimates for a 60 percent cut in all tariffs for the eighteen major industrial countries indicate a global welfare gain of only $1 billion in 1971 dollars, including the welfare gains from substitution among suppliers as the tariff wall declines around countries inside the European Common Market and the European Free Trade Association (Cline, N. Kawanabe, T. O. M. Kronsjö and T. Williams).

The conventional way to derive more impressive results is to appeal to other much more important, usually dynamic, gains. These gains have generally not been quantified, however, and qualitative statements about them are usually unpersuasive to those opposed to liberalization. Thus, for policy as well as scientific reasons, a high priority should be assigned to the development of methodologies for quantifying these gains.

The exploitation of economies of scale should be relatively easy to incorporate, since production function estimates of returns to scale by sector can be combined with predictions of changes in sectoral output levels expected from liberalization. Tests are also needed of the proposition that dynamic gains are derived from increasing efficiency over time in those sectors exposed to freer imports. This could be done by using sources of growth analysis, relating the sectoral residual of technological change to the level of, or change in, protection by sector.

Another approach is indicated by the literature on $X$-efficiency (Harvey Leibenstein). It suggests that, as tariffs fall, inefficiency will be squeezed out of domestic production because entrepreneurs will react to increased pressure from imports by cutting costs. This adjustment might occur even with a competitive domestic market structure; for example, atomistic producers who formerly refused to adopt a technology with high risk but low cost would be forced to do so. The $X$-efficiency approach is subject to measurement (Joel Bergsman), but only under quite arbitrary assumptions. In addition, it cannot convincingly explain why the increases in efficiency do not occur even without liberalization as the result of profit maximization by domestic firms, at least in the competitive case. More generally, the method seems likely to lead to an overstatement of prospective gains from liberalization because in actuality there may be no further room for cost reduction by domestic firms as tariffs fall. Hence it may be less promising than alternative approaches.

A separate but related nontraditional welfare effect is that from increased competition as domestic oligopoly power is reduced by trade liberalization. However, despite the literature on monopolistic competition and the special literature on its response to imports (Richard E. Caves), there exists no specific methodology for quantifying it. There is no method which specifies the correspondence between a percentage reduction of tariffs and the resulting reduction in the degree of concentration (as measured, for

example, by the Herfindahl index), nor any corresponding method for translating these reductions in concentration measures into welfare gains.

Finally, the antinflationary role played by liberalization through all of these effects permits more expansionary domestic policies, and thus creates potential welfare gains through a macro output effect as well as a micro efficiency effect. And analysis is needed of the (largely political) assertion that trade policy itself is unstable, so that the gains from liberalization thus include avoiding the costs of increased trade barriers.

## IV. Trade and Payments Balances

Another area requiring research is the welfare effect of trade balance changes under alternative payments situations. At the policy level, standard trade theory is perceived to have postulated, since Adam Smith and David Ricardo, that any concern about changes in the trade balance is mercantilistic nonsense. Not surprisingly, policy makers reject this absolutist position and pursue their own *ad hoc* assessments of the issue.

Yet the body of trade theory, albeit at a somewhat abstract level, does point to welfare loss from a trade balance deterioration (which can be either exogenous or induced by policy) under a situation of payments deficit. If domestic restraint is employed to offset such a decline, there can be a sacrifice of potential output (James Meade, pp. 171–77). If the exchange rate is depreciated, there will normally be a deterioration in the terms of trade and therefore a welfare loss attributable to a higher real price paid for imports. (This second conclusion, a standard assumption of earlier trade theory, holds even taking account of non-traded goods (Meade, pp. 235–47) so long as supply elasticities of exports and imports remain reasonably high. In more recent formulations the conclusion depends on home and rest-of-world import propensities, in analysis of the transfer problem, but again the "convention" is that a transfer—and therefore depreciation—worsens the terms of trade (Harry G. Johnson, p. 514). Under a situation of payments surplus, on the other hand, a trade balance decrease can mean a welfare gain by providing more goods and dampening domestic inflation.

What is needed is theoretical and particularly empirical work to translate these effects into quantitative measures specifying the welfare change per unit of trade balance change under alternative conditions. In the absence of such guidelines, policy makers facing balance of payments constraints tend to seek trade balance improvements through the use of policy instruments with frequently high costs—export credit subsidization, tax incentives for exports, high preference margins for government procurement from domestic suppliers, efforts to achieve trade balance gains in "reciprocal" trade negotiations, and so forth. The incorporation of explicit measurement of any benefits to be gained from such instruments most probably would have the effect of tempering the excesses of these policies.

Although the move to flexible exchange rates may reduce the urgency of developing such guidelines, they nonetheless remain important. Intervention in the exchange markets persists. And, even with freely floating rates, policy makers should be aware of the welfare effects of specific policy changes via their influence on the exchange rate.

One method of specifying a measure for these welfare effects is to estimate a shadow, or accounting, price for foreign exchange. Given this price, a given change in trade balance will generate a welfare effect equal to the magnitude of the change, multiplied by the ratio of the

shadow rate to the market rate, less unity. For countries in chronic payments deficit, the shadow rate would be above the market rate; for those in surplus, below it. The tradition of purchasing power parity analysis is insufficient for identifying the shadow price, since only by chance will the value of the relaxation of a foreign exchange constraint today be equal to the premium over the market rate implied by a historical tracing of purchasing power parity movements.

For developing countries, there is a substantial literature on estimation of the shadow price (E. Bacha and Lance Taylor). In these countries, where the balance of payments constraint is the most severe, shadow pricing of foreign exchange can reverse some standard conclusions of trade theory. For example, since the formation of a customs union improves the members' joint trade balance by the amount of trade diversion, the normally negative effect of trade diversion can generate positive welfare gains—if the scarcity premium for foreign exchange exceeds the excess cost for purchase of partners' goods rather than outside supply (Cline 1975). (Moreover, since in practice the alternative to integration will be the maintenance of tariffs against partners as well as the rest of the world, this payments improvement will constitute a second-best gain even though elimination of tariffs and their replacement by subsidy incentives might be preferable from an efficiency standpoint.)

### V. Other Issues

The increases in economic interdependence raise a host of other issues for research, many of which have been outlined elsewhere (Bergsten 1973, Peter B. Kenen). One central issue related to welfare gains from trade is the optimal degree of openness of a national economy after taking into account the risks involved in international interdependence. The case for European and Japanese protection and stimulation of their domestic agriculture seems much sounder after their several recent experiences with *U.S.* controls, and threats of controls, on exports of grain and soybeans. The case for a laissez-faire *U.S.* policy toward energy imports appears weak after the actions of the Organization of Petroleum Exporting Countries (*OPEC*). Major savings might be available if there were a body of theoretical analysis on the optimal degree of intervention.

Research is needed on several aspects of the calls for a new international economic order. One issue is the premise that the increase in the oil price benefited the world's underprivileged, which is an important basis for the alliance between *OPEC* and the rest of the developing countries. It seems just as likely that the rise had a highly regressive distributional effect, after accounting for the adverse effects on nonoil less developed countries (*LDCs*) and the poor in industrial countries, and the concentration of oil wealth inside *OPEC* countries. The various schemes of commodity price indexation require analysis to determine the effects on supply and demand, income distribution and worldwide inflation (as indexed *LDC* commodities entered input-output relationships in industrial countries, perhaps with profit margins added at each stage of the transaction). The theory of nonrenewable resource pricing should also be integrated into the analysis of commodity price arrangements.

One proposal in the schemes for a new order, the link between creation of international reserves and foreign aid, would have to be reformulated to focus on the poorest countries in order to be an equitable aid instrument. In any event, the mechanism is unlikely to generate much aid since *SDRs* are unlikely to be created in the next few years (Cline 1976).

The development of theoretical tools for identifying sectors with future comparative advantage is also important in assessing a new economic order, with its tenet of a rising industrial role for *LDCs* in the world division of labor. The importance of the issue is highlighted by the possibility of reductions in the industrial countries' tariff escalation, through the *MTN*. The methodology could also be applied in industrial countries applying selective industrial aids including the United Kingdom.

REFERENCES

**E. Bacha and L. Taylor,** "Foreign Exchange Shadow Prices: A Critical Review of Current Theories," *Quart. J. Econ.,* May 1971, *85,* 197–224

**R. J. Barnet and R. E. Müller,** *Global Reach,* New York 1974.

**J. Bergsman,** "Commercial Policy, Allocative Efficiency, and 'X-Efficiency'," *Quart. J. Econ.,* Aug. 1974, *88,* 409–33.

**C. F. Bergsten,** *Completing the GATT: Toward New International Rules to Govern Export Controls,* Washington 1974.

———, *The Dilemmas of the Dollar: The Economics and Politics of the International Monetary Policy of the United States,* New York 1976.

———, *The Future of the International Economic Order: An Agenda for Research,* Lexington 1973.

———, **T. Horst and T. H. Moran,** *American Multinationals and American Interests,* Washington, forthcoming.

**R. Berner, P. Clark, J. Enzler and B. Lowrey,** "International Sources of Domestic Inflation," in *Studies in Price Stability and Economic Growth,* Joint Economic Committee, Aug. 1975.

**R. E. Caves,** "International Trade, International Investment, and Imperfect Markets," *Special Papers in International Economics, 10,* Nov. 1974, Princeton Univ.

**W. R. Cline,** "Benefits and Costs of Economic Integration in Central America," Washington 1975.

———, *International Monetary Reform and the Developing Countries,* Washington 1976.

———, **N. Kawanabe, T. O. M. Kronsjö, T. Williams,** "Prospective Trade Effects of Tariff Reductions in the Multilateral Trade Negotiations," Washington 1975.

**H. G. Johnson,** "The Welfare Costs of Exchange Rate Stabilization," *J. Polit. Econ.,* Oct. 1966, *74,* 512–18.

**P. J. Katzenstein,** "International Interdependence: Some Long-Term Trends and Recent Changes," *Int. Organization,* Autumn 1975, *29,* 1021–34.

**P. B. Kenen** (ed.), *International Trade and Finance: Frontiers for Research,* New York 1975.

**H. Leibenstein,** "Allocative Efficiency vs. 'X-Efficiency'," *Amer. Econ. Rev.,* June 1966, *56,* 392–415.

**J. Meade,** *The Theory of International Economic Policy. Vol. I: The Balance of Payments,* London 1951.

**R. Vernon,** "Critical Choices: The Industrial Structure of the United States and Europe," mimeo 1975.

HARRY G. JOHNSON, University of Chicago: Baldwin's paper is an admirably careful example of professional empirical study of the effects of tariff elimination. The only question it raises is that, whereas theory stresses the "small triangle" welfare gains from tariff reduction, the implied welfare cost in Baldwin's study is the loss of jobs in the affected industries. If this is the relevant measurement, theory needs reshaping. Apart from that, in order to determine whether costs or gains are "large" or "small," we need a standard of comparison. It seems to me that a change requiring an exchange rate change of $\pm\frac{1}{2}$ percent, or the loss of something over a thousand jobs in only a few states over a ten-year period, is scarcely worth worrying about.

Horst's paper is equally professional. My comments are only that in a time of inflation tax deferral involves considerable real savings to the taxpayer allowed to defer; and that I am unhappy with the work on international companies and "domestic monopoly profit." It seems to me that both measured monopoly profit and competitiveness in production abroad are the results of some kind of technological advantage in the broad sense; and that describing the phenomena of the complex economics of technology as a public good produced by private investment in terms of "barriers to entry" is not helpful to understanding.

Bergsten and Cline promise a research agenda; but what they actually offer is a shopping list prepared by a would-be master chef who is so accustomed to eating in first-class French restaurants that he does not know that established firms like General Mills, H. J. Heinz, Pepperidge Farms, and Birdseye exist to make it unnecessary for the home cook to grind his own corn, bake his own bread, and wash his own potatoes. Their list shows remarkable ignorance of both theory and empirical work in the international economic field.

To illustrate, they ask for work on the external corollary of the Phillips curve. Many economists, basing themselves on "rational expectations," doubt whether the Phillips curve exists domestically; work has been done on the same basis for international inflation; and the phenomenon of devaluation-induced domestic inflation is part of the "monetary approach" contention that devaluation does nothing for the balance of payments in the long run. Dynamic gains from tariff reduction have been dealt with extensively in W. M. Corden's *Trade and Welfare;* the same source deals extensively with the allegedly unexplored problem of oligopoly power. Trade liberalization can only be a one-shot palliative of domestic inflation. If policy makers think that the balance of payments can be improved by trade restriction, especially under floating rates, they are repeating a fallacy exploded by Hume two centuries ago and should so be told. The terms-of-trade effect of devaluation, alleged to be ignored by conventional theory, is well known to be capable of going in either direction. A quarter of a century ago Sidney Alexander showed that if a terms-of-trade effect exists, it should be exploited to the optimum restriction point and devaluation used thereafter; earlier, Meade and others pointed to the correct policy of free trade plus international transfers. Finally, it is odd indeed to find *U.S.* government subsidized sales of wheat to Russia, and the actions of *OPEC,* as proof that the free market cannot be trusted. In short, the state of trade theory is quite adequate to the task the authors want to assign it, even though the relevant theory is not known in Washington.

# The Empirical Evidence on the Monetary Approach to the Balance of Payments and Exchange Rates

*By* STEPHEN P. MAGEE*

This paper is a critical review of the empirical evidence on the operation of the international monetary mechanism under both fixed and floating exchange rates. I limit the discussion to recent studies on the monetary approach to balance of payments (*BOP*) and foreign exchange rate (*FX* rate) determination for individual countries.

The "monetary approach" contrasts with the "component accounts" approach to the *BOP* (see, e.g., Martin Prachowny); according to the latter, the *BOP* is determined by combining behavioral equations for all of the components of the *BOP* with the *BOP* accounts identity. The two approaches can be reconciled in a general equilibrium framework with goods, assets and money. But reduced-form estimates of equilibrium in any one of the three markets should include the exogenous determinants of equilibrium in the other two markets. On this score, some of the monetarists err, just as some trade flow and capital flow empiricists, in estimating only a single structural equation rather than a reduced-form equation.

This error notwithstanding, research efficiency dictates that if we want a simple (Occam's razor) explanation of the *BOP* and *FX* rates, then we should examine international money markets directly rather than the international markets for eggs, potatoes and *AAA* bonds. Critics counter that the monetarists actually apply Occam's scimitar to the research victim and that while they succeed in getting down to bare bones, the skeletal remains sustain insufficient life to be worthy of continued scientific interest. This critique misses the point that the monetary approach is a macro theory and that simplicity is an important criterion for any theory which must be understood and used by policy makers.

A second virtue of the monetary approach is that it avoids the error easily made by component analysts of identifying movements in the *BOP* components with identical movements in the *BOP*. For example, increases in home income and prices reduce trade balances and decreases in interest rates cause capital account deterioration. However, all three of these phenomena lead to increases in the demand for money, and hence improvements in the *BOP* (or appreciation of the *FX* rate) according to the monetarists. We turn to the empirical evidence on these monetary pre-

* Associate Professor, Graduate School of Business, University of Chicago. The author is indebted to the National Science Foundation for research support and to Robert Aliber, John Bilson, Paul Evans, Jacob Frenkel, Harry Johnson, Richard Karplus, Arthur Laffer, Aris Protopapadakis and Myles Wallace for useful discussions. They are absolved from any responsibility for the views expressed and for errors.

dictions. Section I examines the *BOP* under fixed rates and Section II examines *FX* rate determination under floating rates, both via the monetary approach.

## I. The Monetary Approach to the Balance of Payments

The *recent* contributors to the monetary approach to the *BOP* include, inter alia, Robert Mundell (1968, 1971), Harry Johnson (1972), Arthur Laffer (1969), Ryutaro Komiya, Rudiger Dornbusch (1971) and Jacob Frenkel (1971). In the simplest empirical formulation of the theory, the demand for nominal money balances is determined by the price level, real income and the interest rate while the supply of money equals the money multiplier times the supply of high powered money (reserves plus domestic credit). When the domestic supply and demand for money are equated and international reserves are separated from high powered money, the monetary equation can be written as

$$(1) \quad r = a_0 + a_1\Delta \log y + a_2\Delta \log P + a_3\Delta \log i + a_4\Delta \log h + a_5\Delta \log d + e$$

$$r \equiv \frac{R}{R+D}\Delta \log R = \frac{\Delta R}{R+D}$$

$$d \equiv \frac{D}{R+D}\Delta \log D = \frac{\Delta D}{R+D}$$

$R$—reserves $e$—error
$D$—domestic credit $i$—interest rate
$y$—real income $h$—money multiplier
$P$—price level

The elasticities $a_1$, $a_2$ and $a_3$ are identical to those in the domestic demand for money equation. In the discussion which follows, I assume that $h$, the money multiplier, is constant. Ordinary least squares (*OLSQ*) can be applied to (1) if there is no sterilization of reserve changes and if the following conditions hold. Real income is assumed exogenous by the long-run full employment assumption; in the long run, prices and interest rates are also exogenous by the small country assumption. This implies that the demand for money is exogenous for the small open economy and that the domestic monetary authorities do not have control over any real or nominal variables in the economy except the domestic component, $D$, of high-powered money and through equation (1), international reserves. If all of the assumptions hold, $a_5 = -1$, so that a \$1 increase in domestic credit leads, *ceteris paribus*, to a \$1 decline in international reserves.

What happens when this long-run theory is subjected to empirical tests applied to short-run data? The results indicate surprising success empirically, even in the simple formulation just outlined. The best single collection of empirical tests on postwar data are the applications of the model to Australia by Richard Zecher, Sweden by Hans Genberg (1975), Japan by Donna Bean and Spain by Manuel Guitian in Frenkel and Johnson. The money demand variables were significant in quarterly equations for three of the countries: Australia, Sweden and Japan. In all three cases, the real income elasticity of demand for nominal money was positive and significant; the price elasticity was significant and statistically equal to 1 and the interest rate elasticity was negative and significant for Australia and Japan. Direct or indirect versions of equation (1) provided good sample period estimates for $r$, the changes in quarterly reserves as a percent of high-powered money.[1]

What conclusions can we draw from this strong affirmation of the monetarist predictions of the effects of $y$, $P$ and $i$ on the *BOP*? *First*, our initial impulse is to be surprised to get such good results from quarterly data based on a long-run model with so many stringent assumptions. How-

[1] $\bar{R}^2 = .89$ for Australia, .76 for Sweden and .65 for Japan.

ever, the overly strong results were obtained partly *because of* rather than *in spite of* these assumptions. The small country, long-run, full employment assumptions allow empiricists to assume that $y$, $P$ and $i$ are exogenous and unaffected by the supply of money. But if they are not, *OLSQ* estimates of (1) lead to simultaneous equation bias. Consider causation the other way. An exogenous increase in reserves which increases the money supply should increase real income, increase prices and reduce interest rates. Thus, without proper specification and estimation, we don't know whether the coefficients in (1) reflect the demand for money or the effects of the supply of money on $y$, $P$ and $i$. All of these money supply phenomena work in the *same* direction as on the demand side and make single equation estimates of (1) overly favorable to the monetary hypothesis.

*Second,* a \$1 increase in $D$ will not lead to a \$1 decrease in $R$ (i.e., $a_5 = -1$) if some money demand variables are excluded from (1). For example, the increase in $D$ will lead to a depreciation of the (excluded) *FX* rate when the latter is within the band set by the exchange authorities. This absorbs some of the excess supply of dollars so that reserves do not fall by \$1. Thus, *OLSQ* estimates of $|a_5|$ which are significantly less than 1 may imply specification error rather than a refutation of the monetary approach. William Branson uses Pentti Kouri and Michael Porter's capital flow estimates as proxies for $a_5$ and cites their results (of $|a_5|<1$) as "not consistent" with the monetary approach.[2] However, these estimates may have been affected by missing variables (which could explain why they found $|a_5|<1$) and their rudimentary treatment of the current account does not, in my judgment, preclude the possibility that a \$1 increase in domestic credit can also lead to a current account deficit (so that Kouri and Porter's capital account proxies for $|a_5|$ are too low).

*Third,* does a \$1 increase in $D$ lead to a \$1 decrease in $R$ if the central bank sterilizes reserve inflows? The answer is "No." If $a_5=-1$, sterilization will lead to *OLSQ* estimates of $|a_5|>1$. If there is an exogenous increase in the home demand for money, caused, for example, by an increase in world and hence domestic prices, then reserves will flow in. If the monetary authorities sterilize partially, reducing domestic credit by some fraction, $s$, of the reserve inflow, then home money demand is still unsatisfied and more reserves will flow in, etc. Thus, *sterilization leads to a magnification effect on reserves*—and thus on *OLSQ* estimates of the coefficients on the right-hand side of (1)—equal to $1/(1-a_5s)$. Genberg (1975) has handled the problem properly, estimating (1) and a sterilization equation simultaneously for Sweden obtaining $a_5=-1.23$ and $s=-.53$. Since he obtained $a_2=1.06$ (the $t$-values for the three coefficients just listed all exceeded 2) the impact effect on $r$ of a 1 percent increase in $P$ is 1.06 percent, while the ultimate effect, magnified by sterilization, is 3.05 percent. Thus, sterilization causes the monetary predictions to work with a vengeance.

The second point above leads us to expect *OLSQ* estimates of $|a_5|<1$ and the third leads us to expect $|a_5|>1$. Because of these offsetting effects, tests for deviations of $a_5$ from $-1$ using *OLSQ* estimates do not constitute a reliable empirical test of the assumptions underlying the monetary approach to the *BOP*. What does the data show? When Genberg accounted for sterilization in Sweden, he did not find that $a_5$ differed from $-1$; *OLSQ* estimates yielded the same result for Australia while

[2] The proxy estimates for $a_5$ are $-.77$ for Germany, $-.58$ for the Netherlands, $-.43$ for Italy and $-.47$ for Australia. All of the estimates are significantly different from $-1$.

$|a_5|$ was significantly less than 1 for Japan. Guitian's estimates for Spain did not differ from $-1$ but his money demand coefficients were insignificant so that his estimate of $a_5$ is probably dominated by the central bank balance sheet identity. Since none of the *OLSQ* estimates exceeded 1 for the four countries, sterilization apparently never dominated the small country long-run misspecification effect while the latter dominated for Japan.

*Finally*, the monetary approach provides a convenient framework for an analysis of discrete changes in exchange rates in a fixed rate system. Mundell and Dornbusch (1973) emphasize that devaluations are purely monetary phenomena, namely a change in the relative price of two currencies. All of the monetarists stress that devaluations affect only nominal variables and not real variables in the long run; some make the same assertion for the short run as well. Since a devaluation raises the domestic price level, it must raise the nominal demand for money. If domestic credit is constant, the increased demand for money can be satisfied only through increased *FX* reserves generated by a surplus in the *BOP:* in equation (1), $r$ increases by $a_2$ times the devaluation-induced increase in the domestic price level. What is the evidence on the significance of $a_2$ when the change in the exchange rate replaces $\Delta \log P$?

Michael Connolly and Dean Taylor used cross-section data to examine 18 postwar devaluations by small developing economies. They found significant effects of the change in the exchange rate on the change in the *BOP* ($\Delta r$) after 2 years.[3] Genberg (1974) successfully tested the monetary approach on 17 postwar devaluations (approximately one-half were developing countries), obtaining excellent fits.[4] Marc Miles combined cross-section and time series data for 14 devaluations and found the effect of the exchange rate on the balance of payments positive and significant for 5 out of the 14 cases. Thus, the evidence is fairly strong that the monetary approach can be used, with relatively minor modification, to explain the behavior of the *BOP* when exchange rates change.

We turn to the avenue by which devaluations improve the balance of payments: Is it through the trade account, the capital account or some combination of the two? The old elasticities models, the absorption model and the early monetary models which have only goods and money (see Dornbusch, 1973) dictated that the improvement must occur through the trade account. More recent models (Dornbusch 1975; Frenkel and Carlos Rodriguez and Johnson 1975) allow for some combination of excess supplies of goods, debt and equities to add up to the excess demand for money induced by the devaluation. Irrespective of the amount of time required for money and asset markets to adjust, Magee (1974) has shown that the trade statistics will be influenced following devaluations by at least as long as the longest contract (order to delivery) period while some trade adjustment processes will last no longer than this period (Magee 1975b). Finally, Laffer (1976) has argued that devaluations will have no effects on trade. His argument runs as follows. If every asset in the economy (including money) has a corresponding liability (for money these liabilities include government debt) and if the net asset position vis-à-vis foreigners equals zero, then a devaluation-induced price rise will reduce real assets and real liabilities by equal amounts. The two can be restored to their original levels by instantaneously exporting capital and importing money, requiring no change in the trade balance; trade is affected neither in the short nor in the long run. The argument is exotic but it deserves exploration

[3] The $t$-statistic on the $\Delta FX$ equals 2.9; $\Delta FX$ plus the domestic credit variable explains two-thirds of the variation in $\Delta r$.

[4] $\bar{R}^2 = .93$ for $\Delta r$ after 2 years.

since the trade data are so erratic.

Laffer's (1976) study of 15 devaluations found little evidence that devaluations improved trade balances. While these empirical results are subject to alternative interpretations (Magee 1975a), they were strengthened in the study by Miles. He found that, after accounting for changes in both $D$ and government consumption, there was a significant positive effect of devaluation on the trade balance in only 3 out of 14 cases, and it was significantly negative in 2 cases. Hossein Askari, Yves Bizien and Ekram Hossain found that exchange rate changes do affect trade but found empirically that the trade balance effects go both ways.

A modification of Laffer's argument could explain these results. The sum of world assets and liabilities are equal by definition. But some countries' net asset positions are *positive* and others are *negative* vis-à-vis the rest of the world. Trade balance surpluses can follow devaluation in the former case and deficits in the latter.

## II. The Monetary Approach to Exchange Rates

Currently, there are two bodies of literature on the determination of exchange rates. One is an outgrowth of the efficient market analysis of securities markets while the other is the monetary approach. The efficient market approach states that $FX$ markets, like securities markets, are efficient in that no single trader can beat the market in an expected value sense. The evidence examined by Ian Giddy and Gunter Dufey and Richard Levich on floating rates in the recent period is not inconsistent with the efficient market hypothesis. John Bilson and Alan Stockman have shown that forward rates are unbiased estimators of future spot rates. The efficient market approach stresses the absence of unexploited profit opportunities and the importance of market arbitrage relationships. (For interest rate parity see Robert Aliber; for the Fisher effect see Eugene Fama; and for purchasing power parity ($PPP$) see Genberg 1975).

The efficient market approach to $FX$ rates is based on rational microeconomic behavior by individual economic agents. The monetary approach to $FX$ rates is a macro theory which attempts to isolate some of the exogenous determinants of equilibria in national money markets and then explains $FX$ rates as the prices which generate simultaneous equilibria among national money markets Exchange rates and other endogenous variables adjust until the existing *stocks* of each currency are willingly held. I relate here the simplest version of the model. Consider money demand in two countries, $i=1, 2$, and assume that the coefficients of $y$, $P$ and $r$ are the same in both countries. Equilibrium in each market implies the existing stock, $M_i$, is willingly held:

$$(2) \qquad M_i = P_i^{c_1} y_i^{c_2} r_i^{c_3} \qquad i = 1, 2$$

$M_i$ = money stock $\quad P_i$ = price level

$y_i$ = real income $\quad r_i$ = interest rate

$c_1, c_2 > 0 \quad c_3 < 0$

If $PPP$ holds, we can write

$$(3) \qquad P_1 = FX \cdot P_2$$

where $FX$ is the foreign exchange rate. Taking logs and letting $c_1 = 1$, we can write the $FX$ rate equation as

$$(4) \qquad \log FX = \log (M_1/M_2) - c_2 \log (y_1/y_2) - c_3 \log (r_1/r_2) + e$$

Equation (4) was successfully estimated for the German hyperinflation by Frenkel (1975),[5] for the $UK$-German exchange rate by Bilson[6] and for four rates vis-à-vis the

[5] Estimates of (4) yielded $\bar{R}^2=.995$ with $c_1=1.0$ and $-1>c_3>0$.

[6] Estimates of (4) from 1968 through 1975 with the $\bar{R}^2=.89$ and significance of all variables except $c_3$.

dollar by Richard Karplus.[7] All three authors estimated monthly equations. The authors found that the coefficient on *log* $M_1/M_2$ were all greater than 1, but the differences from 1 were not statistically significant. Karplus found equivalence of $c_2$ across the 4 countries. This is a virtue since it is necessary for cross-rate consistency.

Several comments are in order. *First*, authors of the recent empirical studies (Bilson, Stockman, Karplus, et al.) are taking great (and in my view, necessary) pains to make their models consistent with market arbitrage relationships and efficient market behavior. These consistency tests are important if we are to avoid macro model *ad hocery*. *Second*, users of these models should not be led into believing that they will beat forward market estimates of future spot rates since these models, if correct will constitute only a subset of the information incorporated in the market price. *Third*, exchange rate changes generate capital gains and losses (expressed in domestic currency) on central bank holdings of *FX* reserves. Central bank behavior may be affected and the mere statistical conversion of reserves into domestic currency units may cause data problems. *Finally*, both monetary approaches (fixed and flexible) identify the stock demands and supplies for each currency exclusively with the domestic holdings of each currency. This assumption is obviously violated by the presence of external currency markets such as the Eurocurrency market.

I conclude the paper with the interesting attempt by Lance Girton and Don Roper to combine the monetary approaches to explain both fixed and floating rate systems. In effect, they separate the international reserves component from $M_1$ in (4) and place reserves for country 1 on the left side of the equation along with the exchange rate. The composite dependent variable is innovative conceptually since it allows disequilibria among national money markets to be resolved by reserve flows, changes in the exchange rate or some combination of the two. Annual estimates for Canada vis-à-vis the United States for the composite dependent variable yielded good fits. Nicholas Sargen extended the approach with success to Japan and the United Kingdom, with less success to Germany and with poor results for Australia. The fits did not deteriorate in the recent floating rate period for Australia, Canada, the United Kingdom, but they did for Germany and Japan (the countries with the largest parity changes). The vice of the approach is that they may be attempting to explain too much with this single equation. The virtues of the approach are that fixed exchange rate regime coefficients of money demand can be carried over to floating rate periods and vice versa; in the absence of structural shifts, three systems (fixed, floating and managed floating) can be explained using a single equation. Occam would be pleased.

[7] The estimation period was 1973 through 1975 with the $R^2$ (and standard errors) as follows: $R^2=.69$ (.034) for Germany; .84 (.031) for France; .74 (.027) for Italy; and .71 (.030) for Norway.

REFERENCES

**R. Z. Aliber,** "The Interest Rate Parity Theorem: A Reinterpretation." *J. Polit. Econ.*, Nov. 1973, *81*, 1451–59.

**H. Askari, Y. Bizien and E. Hossain,** "The Success of Some Devaluations and Revaluations in the Post-War Era," *Economia Internazionale*, Aug. 1973, *26*, 3–23.

**D. Bean,** "International Reserve Flows and Money Market Equilibrium, The Japanese Case," in *The Monetary Approach to the Balance of Payments*, London 1975.

**J. F. O. Bilson,** "Rational Expectations and the Exchange Rates: Theory and Estimation," manuscript, Northwestern Univ., Dec. 1975.

**W. H. Branson,** "Monetarist and Keynesian

Models of the Transmission of Inflation," *Amer. Econ. Rev. Proc.*, May 1975, *65*, 115–19.

M. **Connolly** and **D. Taylor**, "Testing The Monetary Approach to Devaluation In Developing Countries," *J. Polit. Econ.*, 1976, forthcoming.

**R. Dornbusch**, "Notes on Growth and the Balance of Payments." *Can. J. Econ.*, Aug. 1971, *4*, 289–395.

———, "Devaluation, Money and Nontraded Goods." *Amer. Econ. Rev.*, Dec. 1973, *63*, 871–80.

———, "A Portfolio Balance Model of the Open Economy," *J. Monetary Econ.*, Jan. 1975, *1*, 3–19.

**E. F. Fama**, "Short-Term Interest Rates as Predictors of Inflation," *Amer. Econ. Rev.*, June 1975, *65*, 269–82.

**J. Frenkel**, "A Theory of Money, Trade and the Balance of Payments in a Model of Accumulation," *J. Int. Econ.*, May 1971, *1*, 159–87.

———, "A Monetary Approach to the Exchange Rate: Doctrinal Aspects and Empirical Evidence," *Scandinavian J. Econ.*, 1976, forthcoming.

——— **and H. G. Johnson**, *The Monetary Approach to the Balance of Payments*, London 1975.

——— **and C. Rodriguez**, "Portfolio Equilibrium and the Balance of Payments: A Monetary Approach," *Amer. Econ. Rev.*, Sept. 1975, *65*, 674–88.

**H. Genberg**, "An Empirical Comparison of Alternative Models of Currency Devaluations." manuscript, Grad. Inst. of Internat. Studies, Geneva 1974.

———, "Aspects of the Monetary Approach to Balance-of-Payments Theory, An Empirical Study of Sweden," in *The Monetary Approach to the Balance of Payments*, London 1975.

**I. Giddy and G. Dufey**, "The Random Behavior of Flexible Exchange Rates: Implications for Forecasting." *J. Int. Bus. Studies*, Spring 1975, *6*, 1–32.

**L. Girton and D. Roper**, "A Monetary Model of Fixed and Flexible Exchange Rates Applied to the Post-War Canadian Experience," *Amer. Econ. Rev.*, forthcoming.

**M. Guitian**, "The Balance of Payments as a Monetary Phenomenon, Empirical Evidence, Spain 1955–71," in *The Monetary Approach to the Balance of Payments*, London 1975.

**H. G. Johnson**, "The Monetary Approach to Balance of Payments Theory," in *Further Essays in Monetary Economics*, London 1972, 229–49.

———, "The Monetary Approach to Balance-of-Payments Theory: A Diagrammatic Analysis," *Manch. School*, Sept. 1975, *43*, 220–74.

**R. Karplus**, "The Monetary Approach to Foreign Exchange Rates: The Current Experience," term paper, Grad. School of Business (*GSB*), Univ. of Chicago, Nov. 1975.

**R. Komiya**, "Economic Growth and the Balance of Payments," *J. Polit. Econ.*, Jan. 1969, *77*, 35–48.

**P. J. K. Kouri and M. G. Porter**, "International Capital Flows and Portfolio Equilibrium," *J. Polit. Econ.*, May 1974, *82*, 443–67.

**A. B. Laffer**, "The U.S. Balance of Payments —A Financial Center View," *J. Law Contemp. Probs.*, 1969, *34*, 33–46.

———, "Exchange Rates, The Terms of Trade and The Trade Balance," in T. Willett, ed., *Effects of Exchange Rate Changes*, 1976, forthcoming.

**R. Levich**, "Tests of Foreign Exchange Forecasting Models and Market Efficiency," New York Univ., Bus. School, working pap. 75–88, Nov. 1975.

**S. P. Magee**, "U.S. Import Prices in the Currency Contract Period," *Brookings Papers*, No. 1, 1974, 117–64.

———, "Prices, Incomes and Foreign Trade," in P. Kenen, ed., *International Trade and Finance: Frontiers for Research*, Cambridge 1975a, 175–252.

———, "A Contractual Approach to International Trade and Adjustment," manuscript, *GSB*, Univ. of Chicago, Dec. 1975b.

**M. Miles**, "Devaluation, The Trade Balance, and The Balance of Payments," Ph.D. dissertation, Univ. of Chicago 1975.

**R. A. Mundell**, *International Economics*, New York 1968.

———, *Monetary Theory*, Goodyear 1971.

M. F. J. Prachowny, *A Structural Model of the U.S. Balance of Payments*, Amsterdam 1969.

N. Sargen, "An Empirical Analysis of a Monetary Model of Exchange Rate and Reserve Fluctuations," manuscript, Fed. Res. Bank of San Francisco, Nov. 1975.

A. C. Stockman, "The Market for Foreign Exchange," manuscript, Univ. of Chicago, Nov. 1975.

R. Zecher, "Monetary Equilibrium and International Reserve Flows in Australia," in *The Monetary Approach to the Balance of Payments*, London 1975.

# The Less Developed Countries and the International Monetary Mechanism

By JOHN A. HOLSEN AND JEAN L. WAELBROECK*

The paper discusses the growing involvement of developing countries in the international monetary system.[1] The interest of such a paper would have been slight five years ago. In recent years, developing countries have become large borrowers on the Eurocurrency market; in 1975 they were beneficiaries of over a third of publicized loans. This trend was paralleled by the creation of new International Monetary Fund (*IMF*) facilities adapted to the special needs of these countries' economies.

This paper traces these developments on the basis of a consolidated balance of payments statement for the developing countries which has been synthesized from a variety of sources. The most important are the Direction of Trade statistics, the World Bank's debtor reporting system and the *IMF's* reports on its own operations and on international reserves. Supplemented by Organisation for Economic Co-Operation and Development (*OECD*) information on official grants and private direct investment, estimates of non-trade current account items, and estimates of recent flows from Organization of Petroleum Exporting Countries (*OPEC*) sources, these data of diverse origin add up to a set of accounts which very nearly balances. The residual item—"capital flows *NEI* (including errors and omissions)"—is small, particularly once one accounts for short-term finance in the form of customary lags in payments for imports. (See Table 1.)

The first conclusion is that over the last five years a substantial number of developing countries have climbed into the crowded boat of the international monetary system, which we are trying not to rock. The presence of these new passengers may not make the boat easier to steer; but it is in the interest of all to adjust to their presence. This is recognized in representation the developing countries have in the Interim Committee —the ministerial level group considering reforms in the international monetary organization. It is also reflected in the attention being given to compensatory financing arrangements and the United Nations Conference on Trade and Development (*UNCTAD*) proposals for an integrated commodity program. These discussions are not just about how aid can be channelled to the poorer countries, or even just about the "New International Eco-

* Members of the staff of the International Bank for Reconstruction and Development (*IBRD*); J. Waelbroeck is on leave from the Université Libre de Bruxelles. They are grateful to Wouter Tims, Jo W. Saxe and Elinor Y. Sachse for perceptive comments. The views are their own and should not be quoted as representing those of the World Bank Group, which takes no responsibility for them. The paper deals with very recent developments on which there is as yet very imperfect information. The authors alone bear responsibility for errors of fact or interpretation.

[1] For shortness' sake we will use the term "developing countries" to refer to the subgroup of such countries which excludes the important petroleum exporters as well as countries in Southern Europe and among the centrally planned economies which may sometimes be included under this heading. More specifically, the countries included are those currently classified in the *IMF/IBRD* Direction of Trade statistics as "less developed countries" excluding the group of 17 "oil exporting countries." We will refer to "lower" and "middle income" countries as separate groups. The dividing line is drawn at a level of $200 per capita *GNP* in 1973.

TABLE 1—CONSOLIDATED BALANCE OF PAYMENTS FOR NONOIL DEVELOPING COUNTRIES
(In millions of U.S. dollars)

| | 1970 | 1971 | 1972 | 1973 | 1974 | 1975[a] |
|---|---|---|---|---|---|---|
| Exports (*FOB*) | 35.3 | 36.7 | 44.9 | 64.2 | 92.0 | 88.3 |
| Imports (*CIF*) | 44.6 | 49.9 | 54.7 | 76.2 | 126.5 | 132.5 |
| Trade deficit | −9.3 | −13.2 | −9.8 | −12.0 | −34.5 | −44.2 |
| Net services and Private transfers | 1.0 | 2.0 | 1.0 | 3.0 | 4.0 | 5.0 |
| Current account balance | −8.3 | −11.2 | −8.8 | −9.0 | −30.5 | −39.2 |
| Direct investment | 1.9 | 2.2 | 2.3 | 4.0 | 4.8 | 5.0 |
| Official grants | 1.9 | 2.3 | 2.8 | 3.0 | 4.2 | 4.0 |
| Net medium and long-term (*M&LT*) public loans | 4.9 | 6.2 | 6.6 | 9.3 | 17.5 | 20.4 |
| (from Banks) | (.2) | (1.0) | (1.3) | (3.7) | (6.7) | (7.2) |
| (from *OPEC* sources) | — | — | — | — | (2.4) | (4.4) |
| (other) | (4.7) | (5.2) | (5.3) | (5.6) | (8.4) | (8.8) |
| Net *IMF* resources | −.4 | .1 | .4 | −.1 | 1.6 | 1.6 |
| (Oil facility) | | | | | (.9) | (1.5) |
| Capital flows *NEI*[b] (including errors and omissions) | 2.5 | 2.2 | 2.2 | .9 | 4.7 | 5.2 |
| Change in reserves (+ indicates decrease) | −2.5 | −1.8 | −5.5 | −8.1 | −2.3 | 3.0 |

[a] Estimated on the basis of partial data.
[b] *NEI* = not elsewhere included.

nomic Order" sought by the developing nations; they also correspond to the enlightened self-interest of an international financial community which can only benefit from creating a safety net which reduces the risk of default of large clients.

To be fashionable, the next conclusion is about the new international economic order. Recent developments have demonstrated that through their involvement in the international economic system developing countries have gained a substantial freedom to allow their economies to diverge from world trends. If, as Etienne Kirschen has suggested, economic sovereignty should be defined in terms of power to control a full range of policy instruments, then, by gaining access to borrowing on the international monetary market, middle income developing countries have made a significant step towards the new type of international relations which the third world is demanding today.

It is hardly necessary to stress that these developments also entail risks. They require new debt management skills on the part of the borrowers, as well as willingness on their part to carry out painful adjustments rather than to allow debt to pile upward. An effort is required also of developed countries, from banks which have to weigh the distribution of their risks, and from governments and international institutions, which will inevitably be called upon for help on occasions.

Finally—and this is a conclusion which is never repeated enough—these developments have increased even more the gap between middle and lower income countries. Over the past five years, middle

income countries have achieved per capita rates of growth of 4 percent, while gaining a new freedom to control their economic destiny. During the same years, there has been no growth of income per capita in the lower income countries; especially since the rise in oil prices, these countries have become increasingly dependent on concessionary aid.

### I. Balance of Payments Developments in 1970–75

Table 1 summarizes balance of payments developments in developing countries in 1970–75. By 1975, as the boom was drawing to a close, the middle income developing countries were in a particularly strong balance of payments position. Inflation had reduced the real cost of servicing their foreign debts. A number of developing countries had experimented with more outward oriented trade policies, which had proved very successful in expanding the level of their exports. Many of those countries found that they owned record foreign exchange reserves, which were increasing steadily.

The situation of the lower income developing countries was much less favorable. The most important of these countries are those of the Indian subcontinent, where agriculture had suffered from a series of unfavorable monsoons. Their exports remained very small in comparison to their *GNP* and had expanded rather slowly. As a result the ratio of their foreign debt to export earnings was about twice as high as that of the middle income developing countries. Since a large share of aid to these countries consists of concessionary loans, the debt service to exports ratio was only a third higher than for other developing countries.

Around 1970, developing countries began using the Eurocurrency market as a source of investment credits as well as trade finance. Lending from banks to the developing countries expanded rapidly, growing (on a net disbursements basis) from $200 million in 1970 to $3.7 billion in 1973; by the latter year it accounted for 40 percent of the net disbursements from medium and long-term public loans.

These operations brought developing countries into a new type of relation with the world's financial and banking system. Previously, most of the private capital entering developing countries moved at the initiative of investors in developed countries. Except for trade credits, these capital flows, like aid flows, disbursed quite slowly. Developing countries always had a choice between policies of encouraging or discouraging private capital imports, but these policies operated slowly. Bank loans, because they can be negotiated swiftly and drawn down immediately, and can therefore be used to offset sudden pressures on the balance of payments, fill a gap in the arsenal of policy tools.

Only the middle income countries had access to commercial bank lending. To commercial banks the lower income countries did not appear as particularly attractive risks; the poor countries on their part recognized that they could not afford to bear the cost of borrowing on commercial terms. The export earnings of the poorer countries self-financed—more out of necessity than choice—85 percent of their imports in 1970–73; by comparison, the exports of the middle income countries financed only 80 percent of their imports.

Both groups of developing countries were dealt brutal blows by the rise in oil prices and by the onset of the world recession. The rise in oil prices directly added some $10 billion to import costs in 1974. In absolute terms, middle income countries bore the largest share of the cost. But the burden which the poorer countries had to bear represented a larger

share of their exports. The impact of the recession, by contrast, was most strongly felt by the middle income countries, many of which export minerals and manufactures whose export demand is quite elastic.

The ability of developing countries first to stave off, and later to mitigate the impact of these balance of payments pressures on their economies has surprised observers. One factor which helped finance a high level of imports in 1974 was the large increase in credit that automatically accompanies an expansion of imports; an average payments lag of only one month amounts to more than $4 billion in additional short-term commercial credit when imports grow as rapidly as they did between 1973 and 1974. *OPEC* lending was an important new source of finance which became available during this period. This went both to middle income and to low income countries.

Inflation in 1974 and 1975 has been indirectly helpful. Developing countries are large net borrowers: their foreign debt, net of foreign exchange reserves, was $38 billion at the end of 1973. Every percent of world prices inflation wipes out $380 million of this debt. We assume here that inflation affects only the level of prices, and not the terms of trade of primary products versus manufactures.

It is however the new access to commercial bank borrowing which has made substantial balance of payments deficits possible for the middle and higher income countries. The range of countries which found themselves sufficiently creditworthy to obtain loans from commercial banks is remarkably wide. In 1974 and in the first nine months of 1975, 28 developing countries were involved in the publicized Eurocurrency credits identified in the *IBRD* International Finance Division's quarterly review of the Eurocurrency lending; of these, half may be considered as regular borrowers.

These new members of the international monetary system are accepting substantial risks to maintain the momentum of their growth. Their exports are very sensitive to business fluctuations and their import needs tend to be rigid. They have substantial, fairly short-term foreign debts. Their access to lending of last resort, at the early December 1975 date of writing of the paper, is not proportioned to those risks, though several measures discussed in current negotiations could improve it. The proposed *UNCTAD* integrated commodity program is an example of what is needed. In the *IMF* context, the clauses of the compensatory financing scheme should be modified to enable that scheme to function effectively in a world of inflation. This could provide developing countries with a safety net similar in nature to, but more adequate in size than the *EEC* Stabex scheme agreed on earlier in the year. Developing countries would benefit from any general increase in *IMF* credit tranches. The proposed *IMF* Trust Fund, though it is not designed for short run balance of payments stabilization, would provide *LDC*s with foreign exchange at a critical juncture. Such arrangements would be in the interests of the whole of the international financial community.[2]

The lower income countries, whose *IMF* quotas are very small, have little access to the Fund's regular balance of payments relief lending. An adequate compensatory finance scheme would be helpful to them. The most helpful facility for those countries has been the *IMF* Oil

[2] Added in proof: The access of developing countries to lending of last resort has been improved substantially by the December 24, 1975 reform of the *IMF* Compensatory Finance Scheme, as well as by decisions taken in the January 1976 Jamaica Meeting of the Fund, which increased all credit tranches by 45 percent and organized the Trust Fund. The December 1975 negotiations on the *UNCTAD* integrated commodity programs were unsuccessful, and it does not seem that these negotiations will reach results speedily.

TABLE 2—IMPACT OF BALANCE OF PAYMENTS BORROWING ON THE LEVEL OF *GNP* OF DEVELOPING COUNTRIES AND OF AGGREGATE DEMAND IN DEVELOPING COUNTRIES

| | 1974 | 1975 |
|---|---|---|
| Estimated balance of payments borrowing (billions current dollars) | 7.8 | 11.3 |
| Amount of drop in *GNP* required to close balance of payments in absence of such borrowing | −5.0% | −6.9% |
| Impact on aggregate demand in developed countries | .4% | .6% |

Facility. In 1974 and in the first ten months of 1975, $2.1 billion were loaned to the developing countries under this arrangement. Over half—$1.2 billion—of this went to low income countries.

## II. Deficits and Their Consequences

If external funds had not been obtained rapidly, developing countries would have been caught short by the recession and by the increase in oil prices, and forced to restrict imports drastically. Since in many of these countries nonessential imports are already excluded by strict controls, this would have led to severe disruption of investment and economic activity.

It is not usual in analyzing business cycle developments to think of developing countries as capable of affecting aggregate demand. This is because they are thought of as adjusting their purchases passively to the level permitted by their foreign exchange earnings. This assumption is not correct any more. Less developed countries (*LDC*s) import policies influence demand today in the same way as the *U.S.* deficit of the late 1960's and early 1970's, and the *OPEC* balance of payments surplus. It is worthwhile to evaluate this influence.

The extent to which developing countries would have had to cut back economic activity if they had not had prompt access to additional funds can be estimated using the *IBRD*'s Simlink model. (See Norman L. Hicks.) This model describes the mechanism through which developing countries adjust their imports to the available amount of foreign exchange by changing the level of their *GNP*. We have assumed that in the absence of extraordinary balance of payments financing, foreign exchange reserves of middle and higher income countries would have dropped by $2.0 and $5.0 billion in 1974 and 1975, and that the "normal" capital inflow would have increased from the 1973 level of $17.1 billion with inflation. (See Table 2).

The impact of the *LDC* import surplus on developed countries can be measured using the *IBRD* Multilink system of dynamic multipliers, built up on the basis of simulations of the Link system and of the Desmos *EEC* model. (See Andre Dramais, Waelbroeck and Bert Hickman.) These simulations, which measure the impact on *GNP*s of a change in demand in developed countries, have been used to construct a matrix of demand impacts which is a convenient tool in business cycle analysis. Allocating the *LDC* imports in proportion to developed countries' exports, we see (Table 2) that developing countries have had a perceptible impact on business trends in developed countries. Their balance of payments deficit has sustained demand as much as, say, a vigorous German demand expansion.

The impact on the Eurocurrency market has been particularly important. In 1975, between a third and two-fifths of publicized Eurocurrency credits went to developing countries. The evolution of this key market would obviously have been different if developing countries had not

appeared as creditworthy borrowers. Of course, lifting the veil of institutions, we see that the middle income developing countries have used the Eurocurrency market to borrow a share of the *OPEC* surpluses to finance their oil deficit.

### III. Implications for Reform

In 1975 a number of developing countries have been taking steps to restore balance of payments equilibrium. The *GNP* growth of the middle income countries has slowed down.[3] It nevertheless appears that these countries will have to continue to borrow heavily next year.

That this seems possible is partly due to the exceptionally strong balance of payments and creditworthiness situation of the middle income countries at the turn of the business cycle. The sharp world inflation in 1974 and 1975 has moderated the fall of commodity prices, holding debt service to exports ratios at acceptable levels given historical experience. Exports of manufactures from middle income countries proved surprisingly resilient to the recession.

The fact remains that the middle income group has chosen a course of action which involves risk. The risk is not borne by each country in isolation: any spectacular setback would affect the creditworthiness and growth prospects of the other middle income countries, and would weaken the financial system of developed countries at a very inopportune time.

[3] Thanks to the good monsoon in the Indian subcontinent, growth of the lower income countries has accelerated sharply.

The shock which the latter system would have to absorb would not be very large, however: it could be roughly compared to the impact of serious default by New York City, compounded by the resulting difficulties which would result for other U.S. cities. For the middle income developing countries, the shock would be far greater. These countries' *GNP* is some 7 percent above the level which could be sustained in the absence of balance of payments borrowing, and they would have to resort to very drastic cutbacks in their development programs.

The extent of the risk depends on the pace of recovery, and on the extent to which inflation will continue to whittle down the value of debts as they accumulate. It is clearly desirable to reduce this risk by providing developing countries with a safety net adapted to the special characteristics of their economies. This could be provided in the framework of *IMF*, but other solutions are possible, such as an expanded export stabilization scheme, or the proposed integrated commodity scheme.

### REFERENCES

**A. Dramais** and **J. Waelbroeck**, "A Model of Policy Coordination for the EEC Countries"; and **B. Hickman**, "International Transmission of Economic Fluctuations and Inflation," in A. Ando, R. Herring and R. Marston, eds., *International Aspects of Stabilization Policies,* Federal Reserve Bank of Boston and the International Seminar in Public Economics 1975.

**N. L. Hicks**, "A Model of Trade and Growth for the Developing World," *Eur. Econ. Rev.*, forthcoming.

# How to Fill an Empty Shell

*By* JÜRG NIEHANS*

Two years ago the International Monetary Fund (*IMF*) moved into its new building. It is essentially a shell of offices, 13 stories high, enclosing a glass-domed atrium. The symbolism is hard to resist: In that same year the gold/exchange standard for whose protection the *IMF* had been set up, finally collapsed, leaving the institution as an administrative shell devoid of guiding principles. The basic problem of the international mechanism is how to fill this empty shell.

## I. The Challenge: The Collapse of Commodity Money

The problem arose because of the collapse of commodity money. The gold/exchange standard of Bretton Woods, while providing considerable autonomy for national stabilization policies in the short run, required long-run national price trends in line with production costs in the gold mining industry and thus precluded secular inflation. This constraint was not regarded as a sacrifice imposed on member nations for the sake of international cooperation, but as an expression of the old insight that secular inflation has only costs and no benefits. In particular, the gold constraint was perfectly compatible with the objectives of *U.S.* economic policy as expressed in the Employment Act of 1946 and its subsequent interpretations. As long as these objectives were adhered to, the system could, in principle, last forever. It collapsed essentially because of an abrupt, but lasting change in the course of *U.S.* monetary policy. To identify such a change in a fixed-rate system, the monetary theorist will not examine $M^1$ or the monetary base, since both are determined by the joint monetary policies of all countries and reflect *U.S.* monetary policy with a lag. What he will examine is rather the volume of Federal Reserve credit; this is what *U.S.* policy really controlled. The result, summarized in Figure 1, is remarkable. Up to 1960 the rate of Federal Reserve credit expansion per congressional term was in reasonable conformity with the objective of long-term price stability, generously interpreted; both World War II and the Korean War were followed by contractions, and in the second half of the 1950's the expansion rates were moderate and fairly constant. In 1961, however, the rate of credit expansion increased about sevenfold almost overnight. The shift was not only large and abrupt, but also lasting in nature, and the inflationary course set in 1961 remained unchanged for the next 14 years under three administrations.[1]

What was the reason behind the shift in *U.S.* policy? I believe the main reason was a deficiency of the mainstream economics of 15 years ago. Most of the highly regarded economic theorists shared the notion, based on the Phillips curve, that a lasting reduction in unemployment could be bought at the price of moderate and nonaccelerating inflation. Today we realize, if we did not realize it then, that this

* Abram G. Hutzler Professor in Political Economy, The Johns Hopkins University.

[1] Figure 1 also shows that the annual compound rate of growth of Federal Reserve credit over twelve-year periods was more than five times as high for 1960–72 as it was for 1948–60. It is interesting to note that $M^1$ lags behind Federal Reserve credit during the 1960's and that the lag disappeared when the fixed-rate system collapsed, just as monetary theory for an open economy would make us expect.

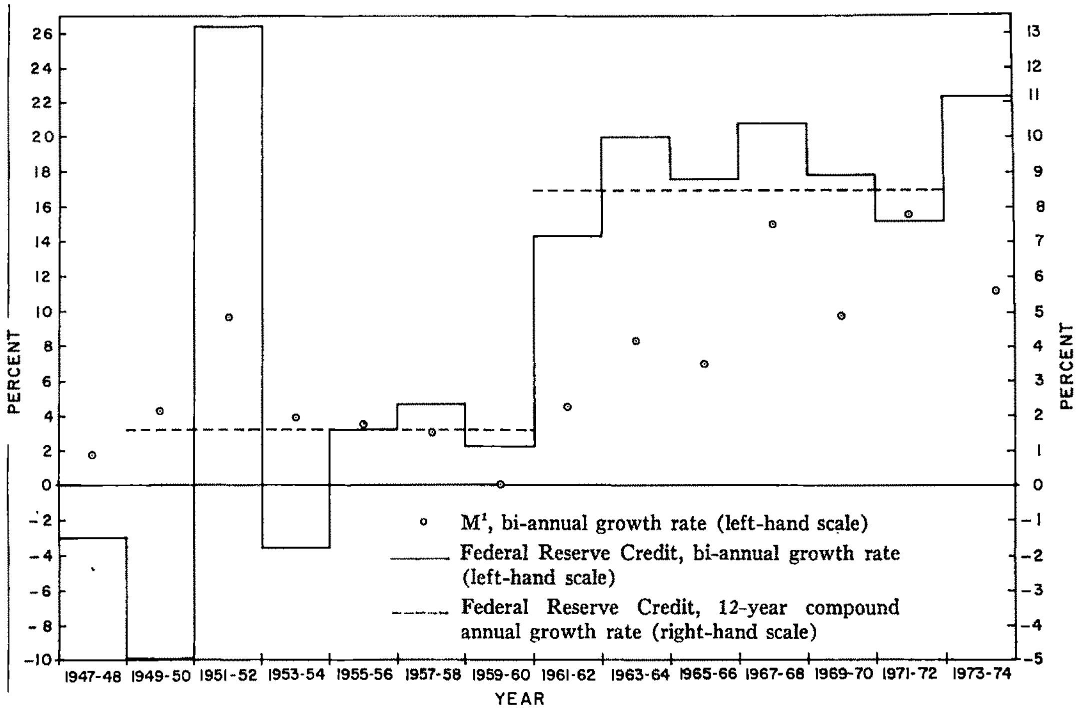



FIGURE 1

notion is, in fact, an illusion. It was under the spell of this "Phillips curve illusion" that the United States changed its monetary course to secular inflation.

Through the fixed-rate system the *U.S.*-generated inflation spread to other nations. If the other countries had consistently used the normal defense of converting dollars into gold, we might still have a viable international system, though probably with a devalued dollar. As it was, most countries, just having recovered from the dollar shortage, eager to show themselves cooperative and suffering from the same Phillips curve illusion, made little use of this defense. As a consequence, the American monetary expansion led to the collapse of the system. This collapse was not the result of technical deficiencies or internal contradictions in the system. It was the result of specific errors in economic judgment at a particular turn of history. Correspondingly, the successful development of new monetary mechanisms will not be a matter of inventing schemes and techniques but of the basic orientation of *U.S.* monetary policy.

## II. Short-Run Response: Adaptation to Inflation

The short-run problem for the international monetary system is to adapt to the inflationary environment created by the Phillips curve illusion. This means, in the first place, that exchange rates between core countries must continue to float and that the efforts to reconstruct a system of "stable but adjustable" rates should not be revived at the present time. These efforts were doomed to failure because central banks which find it impossible to resist the Phillips curve illusion will also find it impossible to submit to an international "harmonization" of monetary

policies. I also believe that in the absence of constraints on the money supply a fixed-rate system is more inflationary than a floating rate system, because it permits each country to externalize a part of the inflationary cost of its own policies which, by a familiar argument, leads to an overproduction of inflation. The immediate reform efforts should thus be concentrated on making floating rates work. To the extent that this is the *U.S.* position, I believe this position is right.

This has a number of implications. First, the process of the demonetization of gold should be completed by allowing central banks to buy and sell gold at competitive market price. It has been argued that a revaluation of central bank gold would be inflationary. I believe this argument is fallacious, because the behavior of central banks depends on what they perceive the true market value to be and not on the artificial numbers used by their bookkeepers. The *U.S.* vendetta against free-market gold serves no constructive purpose and should be abandoned. The monetary role of gold, large or small, will then be decided by market forces.

Second, reformers should stop worrying about the so-called dollar "overhang"; it has been taken care of, not by the oil crisis, but simply by the floating of rates. This can be generalized. There have been many expressions of concern about the lack of control over international liquidity, and ingenious remedies have been proposed. Under floating rates both the concern and the remedies are unnecessary. It is indeed true that there is no longer an international reserve constraint, central banks accumulating foreign exchange reserves at their discretion. It is also true, however, that international reserves have lost their significance for inflation or deflation in the world economy, each country now controlling its own price trend. It can be left to each central bank to determine the proportion of reserves in its total assets.

The third point relates to Special Drawing Rights (*SDR*). With the value of the *SDR* now pegged to a composite bundle of 16 currencies, their interest rate determined by formula and their quantity fixed by administrative fiat, there is no market variable left to adjust to changes in demand and supply of *SDR* relative to the component currencies. This threatens to create a typical Gresham problem of alternating "overhangs" and "shortages": either everybody wants to obtain *SDR,* but nobody is willing to give them up, or everybody wants to pay in *SDR,* but nobody is happy to receive them. To some extent this problem can be suppressed by *IMF* directives, but if it is not solved it will probably prevent the *SDR* from ever becoming a major element of the international monetary system. To solve it, the *IMF* could either decide to buy and sell *SDR* freely against the composite currencies, thus letting the quantity of *SDR* be determined by the market, or permit the value of *SDR* to fluctuate against the composite unit of national currencies. Actually, the 1974 model of *SDR* is devoid of economic significance for the system as a whole, for it is simply a composite of 16 currencies, and reserves in terms of the composite unit can be "homemade" by the 16 central banks exchanging their currencies among themselves. As a consequence, the *SDR* has, in effect, been reduced to an instrument for international transfer payments, and even in this capacity it could just as well be replaced by national currencies.

Under the circumstances, the one important short-run function of the *IMF* seems to be the development, promulgation and enforcement of guidelines for foreign

exchange policies under floating rates. A first set of such rules was, in fact, issued last year. The question is how they might be made more specific. I shall now make six highly tentative and exploratory suggestions on this point.

The first three propositions are of a negative nature. First, I believe there should be no quantitative rules with respect to reserves or foreign exchange transactions at the present time. Trade flows react to exchange rates with a "long and variable" lag (for recent evidence see Helen B. Junz and Rudolf R. Rhomberg), and if, in the case of monetary changes, the eventual working of the purchasing power parity principle gets to be correctly anticipated, they may hardly react at all. In a sense this is desirable because it helps to insulate output and employment against exchange rate fluctuations. At the same time, however, it tends to produce a persistent "overshooting" of these fluctuations accompanied by a large increase in exchange risks (see Rudiger Dornbusch). Central bank intervention to reduce this overshooting, *if based on superior foresight,* might thus, in principle, be beneficial. At the present time, however, reliable criteria for such intervention cannot be established. As a consequence, the *IMF,* while continuing its research on the subject, should refrain from formulating quantitative guidelines, leaving it to the central banks to develop their own criteria and to experiment, individually or jointly, with different intervention systems, perhaps including the crawling peg, dual rates, reference rates or purchasing power formulae.

Second, there should be no quantitative restrictions on the Eurocurrency operations of central banks. The recent performance of the Euromarket has confounded those experts who predicted an imminent crisis, confirming earlier doubts whether they had really a correct understanding of this market. My best judgment is that the Euromarket is a relatively stable segment of the total credit structure, that its liquidity-creating potential, if any, is very low, and that shifts of funds between *U.S.* securities and Eurodeposits, comparable to shifts between different segments of the *U.S.* market, are of minor importance for the international monetary system (see Niehans and John R. Hewson).

Third, there should be no quantitative restrictions on short-term capital flows. Even under the Bretton Woods system speculative capital flows, so often used as a scapegoat for the failures of monetary policy, were not the source of the trouble, but only a symptom of it. Under present conditions they are, in fact, one of the principal stabilizers of exchange rates.

Propositions four to six are positive, but qualitative rather than quantitative. The fourth rule says that central banks, if they wish to create or destroy domestic money by buying or selling foreign assets, should continue to use interest-bearing foreign securities or time deposits rather than foreign cash balances. Since the basic purpose of floating rates is to give each central bank a high degree of monetary autonomy, the influence of country *A* on the money supply of country *B,* while it cannot be completely eliminated, should be reduced to a minimum. This can be achieved by using foreign cash balances only for transactions purposes and using less liquid foreign assets for the accumulation or decumulation of reserves.[2]

My fifth rule is in the same spirit. It says that foreign reserves should be denominated in dollars or in a broad-based composite currency like *SDR.* In both cases the remaining effect of national

[2] This rule would automatically take care of the real problem with the Eurodollar market, which resulted from the large-scale shifts of reserve funds from liquid dollars into less liquid time deposits; if the central banks had shifted into treasury bills, the problem would have been essentially the same.

monetary policy on the rest of the world is spread over a large money market. On the other hand, if "small" currencies like the mark, sterling, or, in the extreme, the Swiss franc are used, there could be a concentrated effect on the German, British or Swiss money market. The United States, in particular, should not undertake large operations in any individual foreign currency, since this might reduce the monetary autonomy of the respective country almost to zero.

The sixth rule, finally, has to do with the relative use of domestic and foreign assets in the conduct of monetary policy. The central bank is confronted with two stabilization problems, the more important one relating to short-run fluctuations in domestic output, the less important one relating to short-run fluctuations in exchange rates. To do its job, the central bank has two instruments, namely operations in domestic assets and operations in foreign assets. In most cases, both instruments will affect domestic demand in the same direction, and the same will be true for exchange rates.[3] The question is how much. It is plausible to conjecture that \$100 million of domestic asset purchases have a relatively stronger effect on domestic demand, while \$100 million of foreign asset purchases have a relatively stronger effect on the exchange rate. If this conjecture were shown to be correct, Robert Mundell's assignment principle could be used to derive the sixth rule. It says that open market operations should be used for domestic stabilization, while foreign exchange operations should be used to stabilize the exchange rate. At the present time, this rule is subject to the qualification that the validity of the underlying conjecture has not been analytically established. The examination of this problem seems to be one of the more urgent small research problems in this field.

I believe that workable guidelines for managed floating can be developed along these lines. To be sure, the result of such a passive adaptation to secular inflation will be no better than second-best. In particular, the hoped for gains in the effectiveness of monetary policy will probably not materialize (see Niehans) and the increased exchange risk will gradually lower the international specialization in production and increase the international diversification of portfolios. However, as long as national money supplies, determined by myopic governments under the spell of the Phillips curve illusion, are not subject to a long-run constraint, this seems to be all the *IMF* can do for the international mechanism.

### III. Long-Run Response: Monetary Constraints

In the long run we should be able to do better. However, real progress cannot be expected from international initiatives, but only from the willingness and ability of governments to constrain the national money supplies over periods of 3–5 years to roughly the rate of real growth. Under one variant, each of several countries, while retaining the flexibility for short-run stabilization, would maintain such a noninflationary trend. Their exchange rates would still fluctuate, but probably not much; speculation would tend to be strongly stabilizing. This could be quite a satisfactory system. Alternatively, the national money constraints could be replaced by an international constraint combined with fixed exchange rates, perhaps with a rather wide band. The *IMF* could then again play its traditional role of strengthening the fixed-parity system

[3] There is some slight ambiguity with respect to purchases (or sales) of foreign securities, since they produce a decline (increase) in foreign interest rates which could result in a capital inflow accompanied by a trade deficit. The latter might conceivably nullify the stimulating effect of the domestic monetary expansion.

by increasing its flexibility through lending. In this case, the national money supplies would deviate from their ideal paths, but again not very much. This is the situation in which the choice between fixed and floating rates, the optimal currency area problem, would finally become a real issue.

The international constraint, if any, could be constructed in different ways. Under an *SDR* standard the amount of *SDR* could be increased by, say, 4 percent a year simply by the payment of interest on outstanding balances; no studies on the "need" for reserves would be necessary. National currencies would then be pegged to *SDR*. While such a system would have some attractive features, it seems unlikely that it can be realized within, say, a decade. Before the *SDR* could serve as the international monetary base, it would have to develop into the principal reserve asset whose total amount has reached a steady state. It is hard to imagine how the *SDR*, created by administrative fiat with little appreciation of market forces, could be brought to this mature state in the foreseeable future.

A second alternative is the return to an international gold standard. While this has the familiar drawbacks, it is the only system of which we know from historical experience that in a noninflationary world it is workable. However, an official return to gold parities by international agreement does not seem likely, because in the absence of secular inflation it offers no significant advantages and in the presence of inflation it is not feasible. I tend to believe, therefore, that a reconstruction of the gold standard, if it should ever take place, would rather begin in the private sector under the pressure of market forces. A distinctive feature of commodity money is its ability to be created and supplied by competitive firms. Indeed, the postwar gold/exchange standard was not an ingenious scheme invented by experts at Bretton Woods, but simply the codification of a system that had evolved spontaneously over centuries. If official monetary policies should further deteriorate rather than improve, gold may play an increasingly important role as a financial asset, as a medium of exchange and a medium of account. After years of secular inflation, governments may then be voted into power which are willing and able to ratify this development by tying their decaying currencies once again to gold. Monetary history would have come full circle. Such perspectives, while vague, should remind the central banks to keep their gold options open. The surest way to prevent the return to gold is to adopt a monetary policy that is not widely different from that under a gold standard.

The third alternative is the dollar standard with the *U.S.* money supply subject to a long-run constraint and other countries pegging their currencies to the dollar. This system would clearly be asymmetrical in the sense that short-run stabilization would be somewhat easier for the *U.S.* than for the other countries. I believe, however, that this drawback would be tolerable and more than balanced by the advantage that this arrangement, with a stable *U.S.* monetary policy, would be eminently workable. In fact, *provided* the core countries bring their money supplies under control, this system is most likely to emerge spontaneously under the pressure of market forces. Probably it is also the system into which the Bretton Woods system would gradually have evolved had the U.S. not changed its monetary course in 1961. In the late 1960's the dollar standard was an American dream, the dream of a bankrupt wildcat banker. For a solid world banker the dream may easily become true. I believe, therefore, that recent *U.S.* efforts to make the international monetary system more symmetrical were shortsighted; like "benign neglect" they were a symptom of

failing leadership. To accept leadership by putting *U.S.* monetary policy back on a noninflationary course, thus encouraging the emergence of a dollar standard, seems to be the most realistic and most promising approach to international monetary reform. The basic problems of the international monetary mechanism are really a problem of *U.S.* monetary policy.

REFERENCES

**R. Dornbusch,** "The Theory of Flexible Exchange Rate Regimes and Macroeconomic Policy," *Scandinavian J. Econ.*, forthcoming.

**H. B. Junz and R. R. Rhomberg,** "Price Competitiveness in Export Trade Among Industrial Countries," *Amer. Econ. Rev.*, May 1973, *63*, 412–18.

**J. Niehans,** "Some Doubts about the Efficacy of Monetary Policy Under Flexible Exchange Rates," *J. Int. Econ.*, 1975, *5*, 276–81.

——— **and J. Hewson,** "The Eurodollar Market and Monetary Theory," *J. Money, Credit, Banking*, Feb. 1975, forthcoming.

HOSSEIN ASKARI, University of Texas: I would like to direct my comments to the paper by Jürg Niehans and in so doing I will have also indirectly addressed the issues presented in the other papers. Niehans is concerned with three related issues—first, the collapse of the Bretton Woods (*BW*) System; second, the short-run management of the current payments system; and third, the long-run reform of the international payments system. Let me take each of these in turn.

Niehans maintains that the *BW* System collapsed because of the lasting expansionary course of American monetary policy which began in 1961. I find this argument either one-sided in its historical perspective or restrictive in its interpretation of *BW*. In the first place, the United States was, for a long time, blamed for the so-called dollar shortage. Then in the 1960's, it was accused of creating a dollar glut. So initially the complaint was one of too strong an American balance of payments position and later it turned to one of too large a deficit. It is clear that if the Europeans were not accumulating enough dollars, they could have devalued with respect to gold and if they were getting too many dollars, they could have revalued. I would, however, agree that the same result could have been achieved if the United States had changed the parity of the dollar to gold; but it chose not to, I think, for good reasons. But, in any case, the rest of the world had a very powerful and direct mechanism for correcting any complaint. Niehans could, however, argue that continuous changes in parity was against the "spirit" of *BW*. I would find such a position as subjective.

In regard to the desired management of the current system, I have two related concerns. First, I think that there should be available a constant purchasing power (in terms of internationally traded goods) and interest bearing Special Drawing Rights (*SDR*). My reason for this is that the oil exporting nations are currently wasting a great deal of resources in haphazard development plans largely because they feel that they have to spend their oil receipts immediately. If, on the other hand, such an *SDR* was available to them, they would be more willing to acquire such an asset in lieu of wasteful development and be possibly motivated to expand output and reduce oil prices. Second, the major balance of payments deficit now resides with nonoil exporting developing nations. But no exchange system can alleviate their deficit position. However, the International Monetary Fund (*IMF*) could, by acquiring *OPEC* balances through the creation of this new *SDR*, make large loans to finance development of such nations. Thus a larger Organization of Petroleum Exporting Countries (*OPEC*) surplus could be indirectly used to finance development of the largest segment of the world population.

In the long run, I agree with Niehans that we should be able to do much better. Any viable long-run system must incorporate some flexibility in allowing for relative autonomy in domestic policy and further not be disruptive in causing unnecessary short-run shifts in resources which are not dictated by long-run scarcities.

Any system of fixed exchange rates will be limiting, as Niehans has pointed out, in flexibility of differing domestic goals. No nation is, I believe, willing to again abdicate control of its domestic monetary tools as required under a system where exchange rates do not change, or at least, do not change frequently.

A system of floating rates, though desirable on many grounds, has two major drawbacks. First, large fluctuations in rates can cause transfers of real resources which may not be desirable from the long-run perspective. Second, and probably much more controversial, floating rates may enhance world wide inflation. I have come to this tentative conclusion from some casual empirical and theoretical considerations.

Economists have, since the writings of Keynes, agreed with the proposition that prices, although flexible upwards, tend to be sticky downward. Furthermore, since the theory of Absorption, the notion that devaluation will, in time, lead to pressures on domestic output, thus increasing prices and leading to an eventual deterioration of the

trade balance is commonly accepted. Therefore, by combining these two pieces of received theory, we can provide a mechanism for world inflation.

A devaluation increases domestic prices. But given that prices are sticky downwards, a revaluation may, on the other hand, only moderate price increases but does *not* lead to price decreases. Clearly if price declines would have occurred in the absence of downward stickiness, then this basic asymmetry becomes crucial. More simply, if a country's exchange rate depreciates by 10 percent, then this will add to inflation by $a$ percentage points; if, at a later date, the exchange appreciates to its old level, then inflationary pressures will be reduced by $b$ percentage points; now because of the basic downward stickiness of prices, $a$ is greater than $b$ and thus the overall impact on inflation is a positive one. In essence, over the time period the exchange rate has not changed (it first depreciated but then it appreciated to its initial value) but because of exchange fluctuations, inflation has increased.

Beginning in August of 1971, many currencies have, with varying degrees of official intervention, floated on world exchange markets. At times, these fluctuations in exchange rates have been *very large*. Thus, these exchange fluctuations, through the mechanism described above, could have contributed substantially to the recent surge in world wide inflation which became substantial beginning in 1971. A casual glance at the data suggests that this ratchet effect of exchange fluctuations could have been the major contributor to world-wide inflation. However, a full scale econometric analysis is needed to assess the significance of this process.

Based on the above casual reasoning and more importantly on ideas expressed in two joint papers with Franco Modigliani, I would recommend a system of sliding parities to provide the best long-run system of international payments.

INDEXATION—MONETARY CORRECTION?

# Government Intermediation in the Indexed Bonds Market

*By* DAVID LEVHARI AND NISSAN LIVIATAN*

Various arguments have been presented in favor of government issuing linked bonds. A well-known argument is that in inflationary periods the government should enable the "small investor" to maintain the real value of his savings. Another argument, which has been made recently, is that the stock of government indexed bonds issued will reduce the government's incentive to use "inflation" as a tax. We do not wish, however, to discuss these arguments, or similar ones, which are based, essentially, on some political or social conceptions. (One could argue as an example that indexing of the government debt makes the inflationary tax a less effective tool and the government may be forced to inflate even more in order to acquire a given amount of real resources from the public.) Instead, we shall concentrate on a class of arguments which stress one major aspect of the problem: that by issuing indexed bonds the government reduces the *inflation risk* faced by the economic agents.

It has been argued that the introduction of a risk-free asset enlarges the feasible set of the private investor and enables him to achieve a better diversification of his portfolio. In addition to this direct contribution to the welfare it is argued that the existence of linked bonds encourages savings under inflationary conditions and reduces the tendency to make excessive investments in real estate and other physical assets. This, it is argued, will reduce the pressure on real resources of the economy and hence reduce the inflationary pressure. Advocates of linked bonds consider these properties to be desirable under inflationary conditions.

In the present paper we take a critical view of these arguments for and against indexed bonds. We shall first argue that it is implicit in the foregoing approach that there is no social economic cost to the government introduction of indexed bonds. It is clear, however, that the government has to finance the interest payments on these bonds. The need to finance these payments imposes a cost on the consumers at large, a cost which may (or is likely to) abolish the benefit of the program. Thus, looking both at the costs and the benefits one cannot argue that the introduction of such bonds necessarily improves consumers' welfare. We shall then consider the arguments that a market for linked bonds tends to encourage savings and reduce "inflationary spending" in the economy. An analysis of this proposition within the framework of a specific model supports the proposition concerning savings but is rather indeterminate concerning the effect on the inflationary process.

If the inflationary process is assumed to be exogenous, we may apply the general principle that *laissez-faire*, perfect competitive markets lead to a Pareto optimal allocation of resources and intervention in that process cannot lead to a superior posi-

* Professors of economics, Hebrew University, Jerusalem, Israel.

tion from the point of view of everybody. In this paper we show in a specific model who benefits and who is hurt from the government intermediation.

Moreover, the inflationary process is at least partially created, influenced and produced by government actions. One has to take into account the ability of governments to manipulate and influence the inflationary process itself and therefore the general principle that competitive markets lead to Pareto optimality cannot be simply applied. One has to take into account the different distributional effects of various government measures. Usage of different rules or acts, like inflationary financing and taxation, will have different distributional implications.

## I. The Formal Model

In order to deal with these problems we need a model for individual and government behavior. For the individual we choose a simple model based on two-period planning.[1]

Consider an individual with a two-period planning horizon who consumes a single commodity. We denote his real consumption of the commodity by $c$ and his initial endowment by $y_1$. We also assume that he has an initial endowment of $\overline{M}$ nominal dollars and wishes to hold $M$ dollars, the real value of which yields direct utility as a consumer's good. He may also hold positive or negative amounts of nonlinked or linked bonds. The dollar values of nonlinked and linked bonds are denoted by $B_N$ and $B_L$ respectively. The *gross* nominal rates of return on a dollar invested in nonlinked and linked bonds are denoted by $i$ and $r(p_2/p_1)$ respectively, with $r$ being one plus the real rate of interest on linked bonds and $i$ is one plus the interest rate.

The individual's budget constraint for the first period, in real terms is given by

$$(1) \qquad y_1 - c + \bar{m} - m = b_N + b_L$$

where

$$\bar{m} = \frac{\overline{M}}{p_1}, \qquad b_N = \frac{B_N}{p_1} \quad \text{and} \quad b_L = \frac{B_L}{p_1},$$

$p_1$ being the dollar price of the commodity in the first period. Denoting by $w$ the real value of the individual's wealth in the second period we have

$$(2) \qquad w = (m + ib_N)\pi + rb_L + y_2$$

where

$$\pi = \frac{p_1}{p_2}.$$

Assuming $p_2$ (next period's price) and hence $\pi$, to be the only random variable in the system we obtain the following expression for the expected value of $w$ (denoted $\bar{w}$) and its variance (denoted by $V_w$):

$$(3) \qquad \begin{aligned} \bar{w} &= (m + ib_N)\bar{\pi} + rb_L + y_2, \\ V_w &= (m + ib_N)^2 V_\pi \end{aligned}$$

where $\bar{\pi}$ and $V_\pi$ denote the expected value and variance of $\pi$ respectively.

The expected utility ($U$) function is assumed to be based on the mean-variance approach for the second period and on separability between periods, i.e.,

$$(4) \qquad U = u(c, m) + f(\bar{w}, V_w).$$

In some of our calculations $f$ is assumed to be quadratic

$$(5) \qquad f(\bar{w}, V_w) = \alpha\bar{w} - \beta(\bar{w}^2 + V_w)$$

with $\alpha$, $\beta > 0$ and $\alpha - 2\beta\bar{w} > 0$. The individual is assumed to maximize $U$ with respect to all the decision variables ($c$, $m$, $b_N$, $b_L$) subject to (1) and (3).

The first order conditions with respect to the three independent variables ($c$, $m$,

[1] The model follows closely N. Liviatan and D. Levhari, "Risk and the Theory of Indexed Bonds," res. report no. 67, Dept. of Economics, the Hebrew Univ. of Jerusalem, Mar. 1975.

$b_L$) can be rearranged to read as follows

$$\frac{u_c}{f_{\bar{w}}} = r \tag{6}$$

$$\frac{u_m}{u_c} = \frac{i}{i-1} \tag{7}$$

$$i\bar{\pi} - r = S(m + ib_N); \tag{8}$$

$$S = \left(-\frac{f_{V_w}}{f_w}\right)V_\pi i2 > 0.$$

The first two tangency conditions have an obvious interpretation. The last condition is $(\partial U/\partial b_L)=0$ and it can be stated alternatively as

$$i\bar{\pi} - r = \left(-\frac{f_{V_w}}{f_{\bar{w}}}\right)\frac{\partial V_w}{\partial b_L} \tag{8'}$$

where

$$\left(-\frac{f_{V_w}}{f_{\bar{w}}}\right)$$

is the marginal rate of substitution between $\bar{w}$ and $V_w$, and is positive for a risk averter $[f_{\bar{w}}>0, f_{V_w}<0]$.

Notice that sign $(i\bar{\pi}-r)=$ sign $(m+ib_N)$. This implies that we cannot have a negative risk premium for nonlinked bonds, $i\bar{\pi}-r<0$, because then $b_N<0$ for everybody and no market equilibrium exists. The same is true for $i\bar{\pi}-r=0$. The only possibility which is consistent with both positive and negative $b_N$ is a *positive* risk premium, i.e., $i\bar{\pi}-r>0$.

## II. The Model of Government Intermediation

We shall assume that the purpose of government intermediation in the bonds market is to increase the size of the market for indexed bonds (or to create such a market when originally it did not exist), through the supply of government indexed bonds. This is based on the presumption that the private sector is reluctant to issue indexed bonds. We assume that in order to finance its indexed borrowing the government lends in the nonindexed market (by buying $b_N$). Thus the government acts as an intermediary in the capital market borrowing indexed and lending nonindexed in order to overcome the private sector's aversion to borrow "indexed." Our purpose is to investigate the implications of this intermediation.

A common argument is that by its intermediation the government enables the individuals (or firms) to reach a better diversification of their portfolios. This would indeed be the case if the individual could ignore the government need to finance any possible losses associated with its intermediation agency. If however we take into account the need of the government to balance its accounts associated with the intermediation, we shall realize that no overall improvement in welfare is possible. To calculate the government "profits" from intermediation, notice that it pays on each dollar it borrows $r\cdot(p_2/p_1)$ $(=r/\pi)$ and gets $i$ on each dollar it lends. Thus in real second period magnitudes the agency's profits are $B^*(i\pi-r)$ where $B^*$ is the value of its borrowing (or lending). Let us assume first that the government balances its accounts by paying out (in the form of transfer payments) to the consumers the agency's profits when the rate of inflation is relatively low $(i\pi-r>0)$ and by taxing the consumers to finance its losses when inflation is relatively high $(i\pi-r<0)$. Thus the government operates on the principle of a balanced budget. (We shall consider later the case where the government does not adhere to this principle.) We shall assume that the amount of transfer payments received by a particular individual is independent of the amounts loaned or borrowed by him. Moreover, we assume that the transfer payments do not vary across individuals so that it is equal to $b^*(i\pi-r)$ where $b^*$ is the real per capita value invested in government bonds. The

individuals are assumed to be aware of the system of transfer payments associated with the intermediation. The individuals' wealth constraint can then be written as:

$$(9)\qquad \begin{aligned} w &= y_2 + (m + ib_N)\pi + rb_L \\ &\quad + b^*(i\pi - r) \\ &= y_2 + [m + i(b_N + b^*)]\pi \\ &\quad + r(b_L - b^*). \end{aligned}$$

The second period variance of wealth:

$$(10)\qquad V_w = [m + i(b_N + b^*)]^2 V_\pi.$$

The term $-rb^*$ is the indexed indebtedness of consumers originating from the expected taxation. This debt is equivalent to any indexed loan of size $b^*$. The assumption is that this debt is no different in any respect from indebtedness through bonds, i.e., the consumers in their behavior take into account all the future tax payments. If there has been no market for indexed bonds without the government's intermediation due to some individual aversion, it is intuitively clear that we could not improve their welfare through the introduction of indexed bonds by the government. The *cost* is the simultaneous introduction of indexed indebtedness of the consumers through their future taxes.

We can now show that government intermediation cannot bring about an overall improvement in the allocation of risk. This can be shown as follows. We see from (9) that the government intermediation is equivalent to a new allocation of $b_N$ and $b_L$ among individuals. Thus if we consider an indexed obligation to pay taxes $rb^*$ as equivalent to indexed borrowing, then we may regard $b_L - b^* = \tilde{b}_L$ as "effective" indexed borrowing. Similarly, we may consider $b_N + b^* = \tilde{b}_N$ as effective nonindexed lending. Summing over all individuals we also have $\sum \tilde{b}_L = \sum \tilde{b}_N = 0$. Hence $\tilde{b}_L$ and $\tilde{b}_N$ represent a *feasible* allocation of loans. If this allocation of loans improves the overall welfare, then the original *laissez-faire* allocation was not Pareto-optimal which seems to be a contradiction. In other words, the government intermediation as described above is just a reshuffle of loans and therefore is unlikely to improve upon the competitive allocation.

We argued elsewhere, using the approach of Kenneth J. Arrow and Robert C. Lind, that there does not seem to be a case for *pooling advantages* resulting from government intermediation in indexed bonds. According to Arrow and Lind, an intermediating company of the kind discussed in our paper will act on the basis of expected values, reducing overall risk if at least two conditions are satisfied. First, the variation in the company's profits must be independent of the other profits or incomes of its shareholders. Second, an individual shareholder's dividend from the company should be made negligible compared with his incomes from other sources, by increasing the number of shareholders. This however does not seem to be the case with the problem in question. Note, first, that the intermediating company's profits will fluctuate with the value of money and so will all the nonlinked incomes and assets of the shareholders. Thus the independence assumption is not met. Furthermore, since everybody is involved in the bond market and everybody pays taxes, we cannot make the share of an individual shareholder negligible compared with his other incomes if the company is to handle a large fraction of the loan market in the economy.

We have assumed so far that the government acts on the basis of a balanced budget. Alternatively, the government could act on the basis of an unbalanced budget financing the agency's losses by printing new money and using its profits to reduce the money supply. Thus when inflation is relatively high so that the agency has losses, the government will increase the money supply and when inflation is rela-

tively low so that the agency has profit, the money supply will contract. It follows that the agency will increase the instability of the money supply and therefore also the value of money. Consequently, some consumers will face a reduction in welfare.

In order to see more accurately the implications of deficit financing of intermediation losses, let us use a simple quantity theory approach. Thus let us assume that next period price level changes are proportional to the quantity of money. The quantity of money per capita in the second period is:

$$\overline{M}_2 = \overline{M}_1 + p_2 g - p_2 b^*(i\pi - r) \tag{11}$$

where $\overline{M}_1$ represents the first period per capita money supply, $g$ represents real government deficit (per capita) through its regular expenditures and $-b^*(i\pi - r)$ represents the real losses per capita which are added to the money supply (as it is financed by deficit). Here inflation is endogenously determined and it is determined by real government deficit spending $g$. The inflation path will be induced by $g$ through its impact on the money supply. If $g$ is assumed to be random so will be the implied price level.

The *quantity theory* assumed states that

$$\frac{\overline{M}_2}{p_2} = k \qquad (k \text{ is a constant}). \tag{12}$$

Combining (11) and (12) and using

$$\pi = \frac{p_1}{p_2},$$

we obtain

$$\pi = \frac{k - rb^* - g}{\bar{m} - ib^*} \tag{13}$$

where

$$\bar{m} = \frac{\overline{M}_1}{p_1}.$$

The expected value and variance of $\pi$ are given by:

$$\bar{\pi} = \frac{k - rb^*}{\bar{m} - ib^*} + \frac{1}{ib^* - \bar{m}} \bar{g} \tag{14}$$

$$V_\pi = \frac{1}{(\bar{m} - ib^*)^2} V_g. \tag{15}$$

We assume that $ib^* < \bar{m}$ which means that the private nonindexed liability to the government in the next period is less than the original money supply. In this case, an increase in $b^*$ leads to an increase in $V_\pi$.

Consider again the "experiment" described earlier where the government intermediation leaves $b_N$ of borrowers constant (shift from private to government borrowing) and reduces lenders' $b_N$ (shift from nonindexed to indexed lending) leaving everyone's $\bar{w}$ unchanged (by setting $r = i\bar{\pi}$). The variance of second wealth is

$$V_w = (m + ib_N)^2 V_\pi.$$

In the case of borrowers, $V_w$ will increase as a result of our increase in $V_\pi$ (through an increase in $b^*$) causing a reduction of welfare. As for lenders ($b_N > 0$) $b_N$ will decline while $V_\pi$ will increase leading to compensated changes or changes in different directions in their welfare. In any case some consumers will be made worse off by the government intermediation.

Of course, in reality the restrictions of the foregoing "experiment" are too tight and the situation is more complicated. For example, the introduction of $b^*$ will affect not only $V_\pi$ but also $\bar{\pi}$ and has additional effects on welfare. If we calculate the difference ($\Delta\bar{\pi}$) between the value of $\bar{\pi}$ with $b^*$ and without, we obtain after some manipulation

$$\Delta\bar{\pi} = \frac{k - rb^* - g}{\bar{m} - ib^*} - \frac{k - g}{\bar{m}} = \frac{b^*(i\bar{\pi}^* - r)}{\bar{m}} \tag{16}$$

where $\bar{\pi}^*$ denotes its value *with* $b^*$. Whatever the sign of $\Delta\bar{\pi}$ (i.e., of $i\bar{\pi} - r$), it will affect different consumers in different

ways. Thus if $\Delta\bar{\pi}$ is positive, it will tend to improve the lenders' welfare and reduce the welfare of large borrowers as it increases the value of their real debts.

In any case since the increase in $V_\pi$ will diminish welfare for a large section of the consumers, we see that a change from the mode of operation of the agency from direct taxation to deficit financing does not change drastically the conclusions.

### III. The Impact of Governmental Intermediation on Consumer's Portfolio and Savings

So far we have considered government intermediation from the point of view of improving the allocation of risk in the economy. There are however other arguments which are in favor of linked bonds. One such argument states that the introduction of governmental linked bonds tends to encourage personal savings in inflationary times. Another argument concerns the *composition* of savings, or of consumers' portfolio. In particular, it is argued that linked bonds tend to channel investment which would otherwise take the form of investment in real assets as a hedge against inflationary uncertainty. Thus the introduction of linked bonds enables the economy to avoid real investment which is directed to the sole purpose of inflationary hedging.

In order to analyze the foregoing arguments, we have to extend our model and include in it a physical asset which may serve as a substitute for linked bonds. We cannot ignore however the fact the investment in real capital has risks of its own. Let us introduce accordingly a physical productive commodity with a random return of the variety of "technological uncertainty." For simplicity, we take this commodity to be identical with the commodity consumed ("wheat"). The quantity of "wheat" held as a productive asset will be denoted by $q$ $(q \geq 0)$ and its gross rate of return (in physical units) will be denoted by $x$. We assume that $x$ is a random variable with a known distribution. The individual's $w$ is then

$$(17) \qquad w_N = (m + ib_N)\pi + qx + rb_L; \qquad b_N = u - c + \bar{m} - m - q - b_L$$

with $\bar{w}$ and $V_w$ given by

$$(18) \qquad \bar{w}_N = (m + ib_N)\bar{\pi} + q\bar{x} + rb_L$$
$$V_{wN} = (m + ib_N)^2 V_\pi + q^2 V_x + 2q(m + ib_N)\,\mathrm{cov}\,(x, \pi).$$

Let us consider now the effect of introducing governmental linked bonds, by means of intermediation, when no such market exists or enlarging the size of an existing market for linked bonds when it is considered to be too small. To analyze the effects of this policy in a framework of general equilibrium seems to be at this stage too presumptuous. We shall confine ourselves instead to a simple experiment on the micro level which may nevertheless provide some clues as to the result of the (missing) market experiment.

The individual experiment is the following. Consider a consumer who reaches his optimum with respect to the variables $c$, $m$, $q$, $b_L$ and $b_N$ at given interest rates and prices ($i$, $r$ and $p_1$) and a given distribution of $\pi$ (this includes the case where no market for $b_L$ exists in which case $b_L \equiv 0$). Let the optimal values of $b_N$ and $b_L$ be denoted by $b_N'$ and $b_L'$. Let us now change the role of $b_N$ and $b_L$ from decision variables to parameters and change arbitrarily the composition of bonds in favor of linked bonds holding their sum constant (i.e., $b_N + b_L = b_N' + b_L'$). Clearly the change imposed on the composition of bonds will require an adjustment in the optimal quantities of $c$, $m$ and $q$. Our purpose is to inquire into the nature of the latter changes which are caused by the increase in the share of linked bonds in the overall portfolio. This experiment may be considered as a method of finding out which of the variables—$c$, $m$ and $q$—are substitutes and which are

complements of $b_L^1$ as the degree of bond linkage is increased. It may therefore provide some indication as to the result of increasing $b_L$ through government intermediation.

The calculations on the effect of increasing $b_L$ (at the expense of $b_N$) on current consumption ($c$) shows that $(dc/db_L)<0$ (as the sign of $r-i\bar{x}$ is negative), so that more indexation in the bond market discourages current consumption and encourages saving. It is somewhat surprising that with all our simplifying assumptions we did not obtain a definite effect on $q$ and $m$. Thus it is not at all clear whether the introduction of linked bonds will reduce investment in physical assets. The reason for this seems to be that not only the real value of money is subject to variation, because of $\pi$, but also the real asset, because of $x$. In a market equilibrium the relative return on these two assets ($u_m$ being the return on money) are adjusted to compensate for the differences in variances and it is not at all clear a priori which of these assets (if any) should be reduced as $b_L$ increases. The only fact which is clear is that $m+q$ will increase since this is the counterpart of the reduction in $c$.

However in the case of $m$ we find that

$$\frac{dm}{db_L} > 0$$

if the condition

$$V_\pi > (\bar{x} - \bar{\pi})\frac{(i\bar{\pi} - r)}{i} \tag{19}$$

is fulfilled.[2]

The results of our experiment seem to indicate the following consequences of introducing linked bonds: a reduction in expenditures on current consumption, an indeterminate effect on investment in physical capital, a decrease in *aggregate* demand for physical commodities (consumption plus capital goods) provided the variance of the value of money is sufficiently large.

A related problem which we analyzed was to compare the effect of an increase in $V_\pi$ in two alternative systems, the first without indexed bonds (nonmixed system) and the second with a "mixed" portfolio including $b_L$.

For the reaction of the mixed and nonmixed regimes to an increase in the variance of the value of money ($V_\pi$) we have the following conclusions. If there is a considerable variability in the returns on physical investment (which is a reasonable assumption), then an increase in $V_\pi$ will increase the demand for current consumption in the nonmixed regime and reduce it in the mixed one. *Demand* for investment in physical assets will increase under both regimes but the increase is likely to be greater under the nonmixed regime. (This is likely to be the case when physical assets and linked bonds are substitutes in the sense defined earlier.) It is not clear whether the current inflationary pressure resulting from an increase in $V_\pi$ will be smaller in the mixed regime. This is related to the fact that there does not seem to be a tendency for a flight from money (and into commodities) under the nonmixed regime.

The main advantages of mixed regimes seem to lie in the tendency of this regime to keep the demand for current consumption and investment in physical capital at relatively lower levels under the conditions specified earlier. If this is considered as an advantage, then it is achieved at the cost of a relatively inefficient allocation of risk in the economy.

[2] A possible intuitive explanation may run as follows. By supplying more "indexing," i.e., by increasing $b_L$ relative to $b_N$, we provide him with more safety as far as changes in $\pi$ are concerned. He may afford, therefore, to enjoy more liquidity in spite of the risk associated with it.

## REFERENCE

K. J. Arrow and R. C. Lind, "Uncertainty and the Evaluation of Public Investment," *Amer. Econ. Rev.*, June 1970, *60*, 364–78.

# Inflation and the Income Tax

By HENRY AARON*

'Tis the Night Before Xmas, and all through the nation
Your bonus means nothing because of inflation.[1]

The recent increase in inflation has elevated the question of whether and how the tax system should be adjusted to cope with rising prices from a matter of academic curiosity to a live political issue. Many other nations, having long experienced high rates of inflation, years ago adopted rules for altering their tax systems automatically as prices rise. Should the United States do so today? If so, what adjustments should be made automatically, and which should be left to *ad hoc* remedy by periodic legislation?[2]

## I. The Problem

Inflation affects tax liabilities in three ways. First, it may alter real factor incomes. Second, it affects the measurement of taxable income. Third, it changes the real value of deductions, exemptions, credits, ceilings and floors, bracket widths, and all other tax provisions legally fixed in nominal terms.

Inflation affects factor incomes by altering such quantities as nominal interest rates, desired money balances, real tax collections (due to inflation-causing issuance of money), and savings. In short, inflation may alter the comparative static general equilibrium and necessitate lengthy adjustments. Whether taxes are collected on nominal incomes, real incomes, or some combination, will affect the properties of the resulting general equilibrium and may have perceptible effects on economic efficiency and on pretax income distribution. Unfortunately, current understanding of these issues is primitive.[3] Whether or not the tax code *should* be written to take account of inflation may well depend on the nature of these effects. However, one must first understand the adjustments necessary to convert the income tax from a tax on nominal income into a tax on real income. In general, the tax code properly disregards the forces that generated income; it is concerned only with measuring correctly the income actually received.

## II. A "Real" Income Tax: Income Definition

All business income may be regarded as the excess of gross receipts (or sales), $S$, over a set of expenses including labor costs —$L$, material costs—$M$, depreciation—$D$, and interest costs, equal to the product of the market interest rate, $r$, times net debt —$B$.[4] Profits, $\pi$, are thus

$$\pi = S - (L + M + D + iB),$$

where $r = i$ the real interest rate in a world of stable prices.

* The University of Maryland and the Brookings Institution. The views expressed are the author's only and do not purport to represent the views of other staff members, officers, or trustees of the Brookings Institution or the Fund for Public Policy Research.

[1] *Mad Comics*, no. 172, Jan. 1975, p. 11.

[2] These questions were examined at a recent conference sponsored by the Brookings Institution and the Fund for Public Policy Research. This paper draws freely from the Conference papers and discussion, which will be published by Brookings in 1976. My acknowledgments to participants are too extensive and too numerous to list, as readers of the forthcoming book will find obvious.

[3] For pioneering efforts see Martin Feldstein (1975b) and Jerry Green and Eytan Sheshinski (1975).

[4] For simplicity, I assume that borrowing and lending rates are identical.

*I define a "real" income tax system to be one that imposes the same real tax burden on any given real transactions regardless of the rate of inflation, $\dot{p}$.*[5] Thus, taxable income from capital must be $\pi(1+\dot{p})$ for any value of $\dot{p}$. In addition, all nominal quantities in the tax code must be inflated by $\dot{p}$, so that their real value is unchanged.[6]

The current tax system clearly does not tax real income. Four changes are necessary. Assume that all prices rise proportionately. Interest rates rise enough to keep real interest payments, measured in current dollars, constant, $i(1+\dot{p})$, and to offset the decline in the nominal value of net debt, $\dot{p}B$, i.e., $r=[i(1+\dot{p})+\dot{p}]$. If the firm uses FIFO accounting, profits, $\pi'$, are given by

$$\begin{aligned}\pi' &= (S-L)(1+\dot{p}) \\ &\quad - \{M+D+(i(1+\dot{p})+\dot{p})B\} \\ &= (S-L-iB)(1+\dot{p}) \\ &\quad - (M+D+\dot{p}B)\end{aligned}$$

Profits are misstated by

$$\pi' - \pi(1+\dot{p}) = \dot{p}(M+D-B)$$

This equation underscores three of the four adjustments necessary to make real taxable income independent of inflation. First, inventory costs must be inflated by the increase in the general price level. Second, depreciation must be similarly adjusted.[7] Third, the firm must be taxed on the capital gain arising from the decline in the real value of net debt, $\dot{p}B$. In that case, profits are equal to

$$\begin{aligned}S(1+\dot{p}) &- L(1+\dot{p}) - D(1+\dot{p}) \\ &- [i(1+\dot{p})+\dot{p}]B + \dot{p}B \\ &= [S-(L+M+D+iB)](1+\dot{p}) \\ &= \pi(1+\dot{p}).\end{aligned}$$

In plain English, a "real" tax system requires that the historical cost of materials and of depreciation be increased by the rate of inflation and that the capital gain on the decline in the real value of net debt be included in taxable income.[8]

The fourth adjustment concerns capital gains. If prices rise 100 $\dot{p}$ percent per year and assets appreciate 100 $r$ percent per year, nominal growth will be 100 $n$ percent ($n=r+\dot{p}$). After t years an asset purchased at 1 will be worth $e^{nt}$, the original purchase price in current dollars will be $e^{\dot{p}t}$, and the real gain in current dollars will be $(e^{nt}-e^{\dot{p}t})$. This amount should be included in taxable income.[9]

Two observations are in order. First, the market interest rate may not rise exactly to $i(1+\dot{p})+\dot{p}$. In that case the real interest cost changes. But many factors in addition to inflation can cause changes in real interest rates. If the objective of the tax system is to impose the same real tax burdens on given real transactions regardless of the price level, then tax burdens should be altered by changes in real interest rates

[5] I assume the initial price level is unity and ignore the choice of an appropriate index with which to deflate the various entries used in calculating income.

[6] An alternative definition of a real tax system would require that real after-tax income be unaffected by inflation. I find this goal overambitious and hard to define. It would require the tax system to offset effects on real factor incomes. And, since the supply of labor might be effected, it would require valuation of leisure.

[7] I assume that the value of depreciable assets is expressed in prices prevailing at the beginning of the accounting period. This assumption is equivalent to assuming that the book value of assets is adjusted annually by $\dot{p}$. Alternative adjustments that defer taxation of gains and losses from changes in relative prices are allowance of *LIFO* inventory accounting and replacement cost depreciation.

[8] John B. Shoven and Jeremy Bulow have argued that if accrued gains on net debt are included in income, the decline in the market value of net debt should also be included. Thus, if bonds decline from 100 to 80 because interest rates rise, a firm should include 20 percent of its net debt in income, since it could repurchase its debt and realize a profit. Although this suggestion has some theoretical appeal, practical objections to it are enormous.

[9] Roger Brinner has shown that if half of capital gains are subject to tax in a world of stable prices, the proportion of capital gains subject to tax when prices are rising is less than half, but approaches half as the holding period increases, i.e., $.5\,(e^{nt}-e^{\dot{p}t})\,/\,(e^{nt}-1)$.

Table 1—Change in Tax Liability as Percent of Before Tax Income if Full Adjustment for Five Percent One-Time Inflation were Introduced

| | First Year and Long-Term Effects[a] | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Manufacturing, Mining Construction | | Transportation, Communication, Electric, Gas Services | | Wholesale and Retail Trade, Services | | Finance, Insurance Real Estate | |
| | First Year | Long Term | First Year | Long Term | First Year | Long Term | First Year | Long Term |
| Change in Taxes | (percent of firms) | | | | | | | |
| *Decline* | | | | | | | | |
| More than 100% | 1.1 | 2.8 | 0 | 4.9 | 0.7 | 2.9 | 6.3 | 7.9 |
| 50–100% | 0.8 | 4.1 | 0.1 | 6.5 | 0.7 | 3.4 | 7.4 | 8.8 |
| 20–50% | 3.7 | 11.7 | 1.6 | 15.8 | 2.9 | 12.0 | 18.2 | 19.9 |
| 10–20% | 4.9 | 16.8 | 1.5 | 16.6 | 4.5 | 16.2 | 13.5 | 13.5 |
| 0–10% | 36.7 | 49.6 | 16.2 | 35.4 | 28.6 | 46.9 | 15.0 | 15.6 |
| *Increase* | | | | | | | | |
| 0–10% | 32.0 | 9.7 | 29.7 | 12.1 | 31.2 | 11.2 | 14.3 | 13.2 |
| 10–20% | 8.5 | 2.0 | 17.1 | 3.4 | 11.9 | 2.9 | 6.0 | 4.7 |
| 20–50% | 6.4 | 1.6 | 17.9 | 2.8 | 11.3 | 2.2 | 7.9 | 6.8 |
| 50–100% | 2.8 | 0.7 | 9.1 | 1.3 | 4.0 | 1.1 | 4.9 | 4.3 |
| More than 100% | 3.3 | 1.0 | 6.8 | 1.1 | 4.4 | 1.4 | 6.4 | 5.3 |

*Source:* T. Nicholas Tideman and Donald Tucker.

[a] Tax liabilities are discounted at 5 percent.

whether caused by inflation, changes in money supply, or "animal spirits."

Second, inflation adjustment may raise or lower tax burdens. The more heavily leveraged a firm is, the more likely is its tax burden to be increased during the onset of inflation by the introduction of inflation adjustments. The size of the debt adjustment depends only on the amount of current inflation. But the size of the adjustment for depreciation depends on the amount of inflation since the oldest depreciable asset was acquired. Thus, the depreciation adjustment begins slowly and, in the face of protracted inflation, only reaches its maximum after many years. T. Nicholas Tideman and Donald Tucker have demonstrated this fact for *U.S.* corporations with assets exceeding $1 million. (See Table 1.) Most firms would pay more tax under a real income tax system than they do under the existing system with the onset of inflation. Large, and in some cases dramatic, tax increases for utilities are the rule because typically they are heavily leveraged. The reverse pattern prevails for financial corporations. Long-term tax burdens are reduced for the overwhelming majority of firms and most dramatically for utilities, a highly capital intensive industrial group. This reversal occurs because heavily leveraged firms, that suffer tax increases from indexation at the onset of inflation, tend to have long livid capital assets and to benefit greatly from revaluation of depreciable assets.

The most striking aspect of Table 1, however, is the extreme variety of impacts of a switch to a real income tax system. A sizeable redistribution of business income tax liabilities would occur.[10]

[10] Sidney Davidson and Roman Weil have calculated the first year effects of these adjustments for the Dow Jones 30 Industrials and 24 Utilities. Their results confirm the wide range of effects. *AT&Ts* taxable income for example, would have risen from $2.490 billion to $4.639 billion in 1974 based on reported transactions. By contrast taxable income of General Motors would have fallen from $1.065 billion to $.254 billion.

### III. A Real Income Tax System: Nominal Quantities

Once income has been defined, actual tax liabilities are calculated by applying tax rates, exemptions, credits, allowances, and deductions, in order to reach taxable income and tax liability. Many of these provisions are expressed in fixed nominal amounts.[11] The real value of those and other magnitudes is reduced when prices rise. Through this channel inflation unambiguously increases income tax liabilities on both individuals and corporations.

According to estimates by Emil H. Sunley and Joseph A. Pechman, the combination of inflation and growth would have increased personal income tax collections from 10.7 to 16.3 percent of adjusted personal income between 1960 and 1975 if Congress had not modified income tax laws. Real growth alone would have increased revenues from 10.7 percent to 12.3 percent of adjusted personal income. In fact, personal income taxes rose from 10.7 to 11.4 percent of adjusted personal income, between 1960 and 1975, indicating that tax changes enacted periodically by Congress counteracted not only the increase in revenues generated by the impact of inflation on the rate structure, but also part of that caused by real economic growth.

The composition of the discretionary tax reductions differed slightly from the reductions that indexation of the rate structure would have generated. In particular, discretionary tax changes have reduced taxes more than indexation would have done for those with real incomes in 1975 dollars below $25,000 and over $200,000, and have reduced them less for those with incomes between $25,000 and $200,000. The difference in disposable income is less than 1 percent for those with incomes of $10,000 to $25,000.

A long-standing argument against indexation of nominal quantities has been that a rate structure fixed in nominal terms acts as a built-in stabilizer. When prices rise, real tax collections increase, dampening aggregate demand and thereby curtailing inflationary forces. This argument is incomplete, however. If inflation and excess unemployment coexist, the reduction in aggregate demand caused by the inflation-generated increase in taxes will tend to stabilize the price level, but aggrevate unemployment.

Based on simulations performed on the SSRC-MIT-PENN model, James Pierce and Jared Enzler report that the impact on real and nominal economic aggregates of an exogenous increase in demand is affected negligibly by indexation of the nominal quantities in the personal income tax.[12]

### IV. What Should Be Done?

> A government's decision as to elimination of each of the two types of inflation effect is logically independent of its decision with respect to the other. It may choose to index the tax system so as to tax income expressed in prices of some former year instead of income in current prices however the latter may be measured; to eliminate distortions in the measurement of income in current prices; to do neither of these things; or to do both of them. [Edward Denison]

Because the choice about what to do involves guesses about the outcome of research not yet performed and political judgments on which we are all amateurs, I shall try to characterize the views expressed at the Brookings Conference and turn to my own only in conclusion.

[11] For example, in 1975 married couples paid 14 percent of the first *$1,000* of taxable income in tax; the personal exemption was *$750;* the per capita credit was *$30;* corporations paid 20 percent on the *first $25,000* of profits, and 22 percent on the *second $25,000*; the low income allowance was *$1,900* and the maximum standard deduction was *$2,600*.

[12] In correspondence, Ray Fair reports similar results for identical simulations in his model.

Few thought that indexation of nominal quantities (i.e., of the rate structure) was intrinsically very important at rates of inflation experienced since 1960 or even since 1970. Congress has shown itself quite capable of reducing taxes sufficiently promptly to keep aggregate collections roughly constant. The difference between discretionary tax cuts and the hypothetical results of an indexed 1960 law may be viewed either as expression of a collective desire to reallocate tax burdens or of irrational Congressional reticence to tamper with the divisive issue of the rate structure. But, in either case, the difference is small. (Actual reductions have concentrated on personal exemptions, the standard deductions, low income allowance, maximum tax on earned income, the personal credit, the earned income credit, and so on.) The differences between the stabilizing properties of an indexed system and those of a periodically amended system are negligible.

In the eyes of those who place primary emphasis on indexation of the measurement of income, indexation of the rate structure was viewed by some as likely to further their primary goal (by alerting policy makers to the issue) and by others as likely to impede their primary goal (by letting the public think that partial action was enough). According to another view, indexation of the rate structure would set back the cause of tax reform. Tax reforms typically increase revenue by widening the base, but Congress seems loathe to enact revenue-increasing tax bills. Consequently, tax reform is most likely to occur when automatic increases in revenues create the opportunity for discretionary reductions. Indexation would drastically reduce the number of such opportunities.

While these judgments are clearly open to challenge, the key point is that no participant at the Brookings Conference argued that indexation of the rate structure is very important and all felt that the issue is primarily political, not economic.

The case for or against the tax base is far more complex and rests to a much greater extent on economic analysis. Unfortunately, much of that analysis has not been done. The basic argument for indexation of the tax base flows naturally from modes of analysis congenial to economists. First, all of the postulates of welfare economics rest on marginal equalities among real quantities. If taxes fall on nominal quantities that are not proportional to real quantities, excess burdens result. Second, a nominal tax system violates principles of horizontal equity and may increase the volatility of interest rates during inflationary periods. Balancing these advantages of a real income tax system are the added compliance costs for business and households. At 30 percent inflation, most would feel that benefits of a real income tax system outweighed added compliance costs; at 1 percent inflation the reverse is true. The practical question, then, is how much inflation makes adjusting the tax base worth the added trouble. If a decision is made to adjust the tax base, how should it be done, and how quickly?

The latter question gains force from the estimates compiled by Tideman and Tucker, summarized in Table 1. They suggest that the redistribution of tax liabilities among and within industries, in both the short and the long run, from even a short burst of inflation would be sizeable. The effects of sustained inflation would be even greater. While the distortions of a nominal tax system have little to recommend them, many may have been capitalized into equity prices. Their removal would create windfall gains and losses, perhaps large ones. This possibility suggests the need to consider how an indexed system should be introduced.[13] This issue was raised, but not

[13] Should it be enacted now, but become effective only after several years? Should taxpayers be required

addressed, at the Brookings Conference.

Furthermore, all taxes create distortions and excess burdens. Under a nominal tax system, inflation changes those distortions. But it is not clear which set of distortions is more serious.[14] Consider the following argument. "In the absence of inflation, income from capital is undertaxed, largely because only half of capital gains is included in income, the interest on state and municipal bonds is exempt, and businesses may claim excessively rapid depreciation (especially on structures, see Paul Taubman and Robert Rasche) and the investment tax credit.[15] Inflation operates through a tax system based on historical cost in the same manner as an imperfect capital levy (see Peter A. Diamond). Accordingly, inflation at least up to some (what?) level offsets existing imperfections in the definition of income." If this view is accepted, then adjustment of the tax base for inflation may be desirable only as part of a general reexamination of the taxation of income from capital.

Some participants in the conference held that indexation should reproduce the real tax burdens in an inflationary world that the *present* tax code would generate if prices were stable. I find this view untenable. First, some existing provisions were introduced or are retained in part to offset inflation. (Abandonment of the reserve-ratio test seems the clearest example.) Second, the view that removing the distortions of inflation deserves attention whether doing so aggravates or ameliorates existing imperfections seems to place form before substance.

## V. Conclusion

The most useful conclusion I can offer to this incomplete treatment of the issues surrounding indexation is an agenda for future research, however partial and idiosyncratic it may be. First, models of the effects of inflation through the tax system on the real equilibrium and on the growth path of the economy are in an early stage of development. Until we have a better understanding of how the failure to index affects real quantities, it is hard to argue persuasively that conversion to a "real" income tax is worth the legislative trouble, the probability that a radically new approach to taxation will entail costly and lengthy shakedown period, and the further erosion of the simple, nominal dollar standard we live by (see Arthur Okun). Although this research is important, it is immensely complicated and so far has required drastic and essentially important abstractions from reality if any analytic results are to be obtained. For this reason, I am pessimistic that it will contribute much *independently* to the debate about indexation.

Second, extensive research on problems of transitional equity is clearly indicated and is likely to contribute significantly to the policy debate. How great are the dislocations of switching from the existing tax system to an indexed one? Who gains? Who loses? How can losers be compensated? Should they be? Clearly these estimates must entail models of the impact of inflation on real quantities under a "nominal" income tax system.

Finally, a whole set of questions relating to administrative costs (a low prestige subject which economists of even moderate ambition discreetly ignore[16]) need to be examined. Is the added cost of an inflation adjusted system big or small? If it is big,

---

to calculate liabilities under both nominal and inflation-adjusted bases, and pay tax on a weighted average, where the weights shift over time? Is some other course preferable? On transitional equity, see Feldstein (1975a).

[14] See Martin Bailey.

[15] This argument presumes that income, rather the consumption, is the proper tax base.

[16] A notable exception, but only at the theoretical level, is Walter Heller and Karl Shell.

are there short cuts to indexation that remove most of the distortions but cause few of the adjustment problems? Precisely what changes must be made to index the tax system?

When the answers to some of these questions are available and when we are better able to tell whether the inflation of the past few years was an unwelcome intruder or the harbinger of an inflationary brave new world, we will be in a better position to determine the desirability of indexation. Until then, I suggest that we recall Mort Sahl's definition of a "modern Republican," a person who is willing to try new policies —but not now, and that we adopt the mantle of "modern indexers."

REFERENCES

**M. Bailey**, "Inflationary Distortions and Taxes," paper presented to the Brookings Conference on Inflation and the Income Tax System, Washington, Oct. 30–31, 1975.

**R. Brinner**, "Inflation and the Definition of Taxable Personal Income," paper presented to the Brookings Conference on Inflation and the Income Tax System, Washington, Oct. 30–31, 1975.

**S. Davidson and R. Weil**, "Inflation Accounting: Some Income Tax Implications of Recent FASB Proposals," presented to the Brookings Conference on Inflation and the Income Tax System, Washington, Oct. 30–31, 1975.

**E. Denison**, "Price Series for Indexation of the Income Tax System," paper presented to the Brookings Conference on Inflation and the Income Tax System, Washington, Oct. 30–31, 1975.

**P. Diamond**, "Incidence of An Interest Income Tax," *J. Econ. Theory*, Sept. 1970, *2*(3), 211–24.

**M. Feldstein**, "On the Theory of Tax Reform," Harvard Institute of Economic Research, disc. pap. 1975a.

———, "Inflation, Income Taxes, and the Rate of Interest: A Theoretical Analysis," Harvard Institute of Economic Research, dis. pap. no. 414, July 1975b.

**J. Green and E. Sheshinski**, "Budget Displacement Effects of Inflationary Finance," (mimeo) Sept. 1975.

**W. Heller and K. Shell**, "On Optimal Taxation with Costly Administration," *Amer. Econ. Rev. Proc.*, May 1974, *64*, 338–45.

**A. Okun**, "Inflation: Its Mechanics and Welfare Costs," *Brookings Papers*, 1975, *2*.

**J. Pierce and J. Enzler**, "The Implication for Economic Stability of Indexing the Personal Income Tax," paper presented to the Brookings Conference on Inflation and the Income Tax System, Washington, Oct. 30–31, 1975.

**J. B. Shoven and J. I. Bulow**, "Inflation Accounting and Corporate Profits," (mimeo).

**E. Sunley and J. Pechman**, "Inflation Adjustment for the Individual Income Tax," paper presented to the Brookings Conference on Inflation and the Income Tax System, Washington, Oct. 30–31, 1975.

**P. Taubman and R. Rasche**, "Economic and Tax Depreciation of Office Buildings," *Nat. Tax J.*, Sept. 1969, *22*, 334–46.

**T. N. Tideman and D. Tucker**, "The Tax Treatment of Business Profits Under Inflationary Conditions," paper presented to the Brookings Conference on Inflation and the Income Tax System, Washington, Oct. 30–31, 1975.

# Indexing and the Capital Markets

By WILLIAM POOLE*

The inflation of the past few years has generated substantial interest among economists in the theory of price level indexed bonds. But this academic interest, as measured by the number of papers on the subject, stands in sharp contrast to the market interest in indexed bonds as measured by the number of new issues of such bonds. Indeed, a recurring theme in recent papers on indexing is the question of why indexed bonds do not exist in the United States.

The analysis of the nonexistence of indexed bonds, and of the potential role of these bonds more generally, may be conveniently divided into four topics, and this paper is organized accordingly. The first topic concerns the analysis of indexed bonds within the context of a total portfolio. The second concerns the role of indexed bonds in changing the economy's response to changing rates of inflation. The third concerns the use of these bonds to eliminate the distortion to the time shape of debt service caused by inflation. The fourth topic—one given insufficient attention in my view—concerns the practical problems of index construction and related matters.

## I. Portfolio-Theoretic Analysis

The recent application by Stanley Fischer (1975a, 1975b) and others of standard portfolio theory to index bonds has substantially improved the analysis of these instruments. It is now generally accepted that, roughly speaking, the household demand for indexed bonds—a demand that may be negative if bonds are issued as a source of funds—depends on how the bonds fit into a total portfolio consisting of human and physical capital, equities, nominal bonds, and nominal debt. Similarly, a firm's supply of indexed bonds—a supply that also can be either positive or negative—depends on how such bonds fit into a capital structure consisting of nominal bonds and equities financing the real and/or financial assets owned by the firm. In the abstract, the same portfolio theory is applied to households, financial firms, and nonfinancial firms and these different types of economic units take on separate theoretical identities according to the assumptions made about the expected yields and the variances and covariances of the yields on the different types of assets and liabilities. Of lesser importance, the different types of economic units may also be assumed to differ in their attitudes toward risk.

Alan Blinder (1975), using a portfolio model, has recently suggested that indexed bonds do not exist because a firm would want to choose as the index the price of its own output while a household would want a general price index. Blinder's ingenious solution to this problem is his proposed "National Inflation Mutual Fund" that would buy bonds indexed to individual industry prices and sell bonds indexed to a general price index. As Blinder emphasizes, however, a firm would want to issue bonds indexed to its own selling price only if that price were positively correlated with the firm's profits. A firm would not want to issue such bonds if cost-push forces frequently forced up its own prices and simultaneously reduced its profits.

Blinder's cost-push argument can be

* Professor of Economics, Brown University.

generalized; I would conjecture that *U.S.* experience has been for inflation surprises to be associated with lower real yields on human and physical capital independently of whether the inflation is higher or lower than anticipated. On the one hand, inflation lower than anticipated has usually been associated with business cycle downturns and, therefore, with reduced returns due to unemployment of labor and capital. On the other hand, inflation higher than anticipated has typically occurred during or just after wars. War absorbs part of the national output and leaves less output available to serve as the real return to labor and capital.

The cost-push argument emphasized by Blinder—applied, for example, to the effects of poor harvests and the Organization of Petroleum Exporting Countries (*OPEC*) cartel on inflation in 1973–74—may or may not provide a satisfactory explanation for the recent inflation; there is still the question of whether the cost-push disturbances would have had much effect in the absence of a prolonged period of above-normal money growth. But whatever the merits of the various theories of inflation, if it is true that unanticipated accelerations of inflation are correlated with reduced supplies of domestically available goods, then it is clear that issuance of indexed bonds, whether linked to general or industry prices, will not be attractive to firms. On this view the standard portfolio models suffer from a symmetry assumption, the assumption being that if equity-type investments have real yields that are affected by inflation, then those yields change in one direction when inflation is higher than anticipated and in the opposite direction when inflation is lower than anticipated.

An observation supporting my conjecture is that long-term corporate bonds almost without exception are callable well before maturity. I have recently collected data on all *A*aa and *A*a long-term (20 years and over) corporate bonds issued in the United States from 1965 to mid-1974; of my sample of 535 issues, only two were noncallable to maturity and none of the others had more than 10 years of call protection. For no bonds were the call premiums very large. Similarly, individual home mortgages typically have very limited call protection. The callable bond and callable mortgage are well-designed instruments for economic units whose real earnings are depressed by inflationary surprises in both directions; real interest costs fall with positive unanticipated inflation but rise only to a limited extent with negative unanticipated inflation because interest rates fall and high nominal interest bonds and mortgages are called and refinanced.

From the recent work it is clear that portfolio theory does not adequately explain the complete nonexistence of indexed bonds in the United States. While it may or may not be true that the representative household and representative firm face parameter values that make indexed bonds unattractive, it must be true that there are many nonrepresentative households and firms for whom indexed bonds would be highly desirable. A diversity of products tailored to different tastes and technology is the rule in competitive markets, and certainly the existing diversity of financial assets suggests that there is no reason why indexed bonds should not have at least some place in the market.

## II. Macroeconomic Performance Under Changing Inflation Rates

It has been argued, especially by Milton Friedman (1974), that widespread indexation would substantially ease the transition to a secularly lower inflation rate and thereby improve the odds that the government will maintain policies consistent with ending inflation. The argument is that

unanticipated declines in inflation cause unemployment because real wages and real debt service rise. Moreover, indexation of the tax structure would eliminate the government's easy access to real revenue through inflation.

Whatever the merits of this position in general, its applicability to the capital markets seems somewhat limited. As noted above, corporate bonds and home mortgages are already rather effectively indexed on the downside through call provisions. The federal government is not especially vulnerable to lower inflation with respect to the real burden of its debt because it has so little long-term debt outstanding and because of its extensive taxing and borrowing capacities. State and local governments are by far the most exposed issuers of long-term debt; municipal bonds are generally not callable and a substantial quantity of debt with distant maturities is outstanding.

### III. The Time Shape of Debt Service

Much recent work on the home mortgage market has emphasized that the traditional fully-amortized fixed rate mortgage carrying a nominal interest rate reflecting inflationary anticipations is a poor instrument for the typical household.[1] The traditional mortgage has a constant nominal debt-service stream and, therefore, a declining real debt-service stream in an inflationary period. If two mortgages, one issued in an inflationary environment and one in a noninflationary environment, have the same anticipated real rate interest, then the one issued in the inflationary environment may involve such rapid repayment in real terms that the typical household may be unable to meet the repayment schedule even though the household's real income prospects are fully adequate to meet the longer real repayment schedule in a noninflationary environment. Other borrowers, such as nonfinancial firms and state and local governments, may face a similar problem.

Many borrowers are able to adjust their borrowing and lending behavior using traditional debt instruments in such a way that the time shape of real debt service of a portfolio package is not changed by inflation. But if real income, real spending flows, and real debt service are to be unaffected by inflation, then during an inflation the portfolio package required to finance a given capital project necessarily involves a larger accumulation of debt, both nominal and real, in the early years of the life of the project; the present value of the distant real debt service must be represented by a currently large value of the principal of nominal debt.

The shortening of the effective maturity of the debt when conventional debt instruments are used in an inflationary period makes all borrowers more vulnerable to bankruptcy. The problem is compounded by the fact that lenders typically do not realize that conventional accounting practices can seriously distort the reported income position of a given borrower. The inflation premium part of nominal interest payments is actually repayment of principal rather than real interest expense.

The use of price level indexed bonds and mortgages would prevent inflation from shortening the effective maturity of debt and would, therefore, reduce the real distortions caused by inflation. In the absence of indexed securities the market has, however, produced some imperfect substitutes for indexed securities, substitutes in the form of adaptations of conventional instruments. The clearest example is the use of the so-called "equity-kicker" clause in commercial mortgages in return for a lower nominal interest

[1] See especially Poole (1972), Donald Lessard and Franco Modigliani (1976).

rate. This provision gives nominal debt service an upward tilt over time, reducing the effect of inflation in speeding up real debt repayment. But market adjustments are surely incomplete; few lenders yet understand the appropriateness in an inflationary period of borrower plans calling for a prolonged rise in the nominal debt outstanding to finance a given capital project.

### IV. Practical Problems with Indexed Bonds

The practical problems of selecting an appropriate index are well-known. These problems, however, are minor compared with the problem that contracts with indexing provisions are much more susceptible to tampering than contracts defined solely in nominal terms. In Israel, many securities issued in the early 1950's were linked to the exchange rate of the dollar, but the indexing provision became worthless due to the government maintaining the official exchange rate at artificially high levels. Later, following linkage to the Israeli Consumer Price Index, there was extensive political controversy over the use of government subsidies to hold down the index and over the construction of the index itself.[2] The index construction problem is much less serious with indexed wages than with indexed bonds because wage contracts are renegotiated at relatively frequent intervals.

The practical problem with indexed securities goes well beyond the possibility of government tampering. The deeper problem is that public attitudes as expressed through democratical political and legal processes provide little protection to indexed contract provisions. Recent experience with wage and price controls in the United States and elsewhere involved governmental cancellation of private contract provisions calling for increases in nominal wages and prices. These cancellations did not meet with any significant opposition; indeed the introduction of controls has generally received overwhelming public support.

At the same time, recent wage-price control experience clearly demonstrates the strength of attitudes involving nominal magnitudes. Freezes have generally commanded widespread support while wage and price rollbacks have received relatively little support. Whatever the reasons for, or rationality of, these public attitudes, they are important and persistent. Given the potential for manipulation of indexes and given public attitudes toward the enforcement of contracts, it is not surprising that indexed bonds have not yet appeared in the United States. No investor can have much confidence that the real yields promised by indexed securities would be realized in the face of a significant escalation of inflation.

### V. Concluding Comment

My guess is that should indexing become popular in the United States the reason will not be the triumph of economists' ideas. Rather, widespread indexing will reflect fundamental changes in the economic environment and in perceptions of the environment. Monetary instability uncorrelated with real shocks will change the variance-covariance properties of different types of assets and make indexed bonds more attractive. With sufficient price level instability, the benefits of indexed bonds will exceed the costs imposed by the practical problems, and public attitudes will force changes in the political environment so that more emphasis will be placed on real magnitudes. Successful innovation of indexed bonds will deserve one cheer—three cheers for indexing plus

[2] See A. Rubner (1960).

two negative cheers for the monetary instability responsible for the event.

## REFERENCES


A. S. **Blinder,** "Indexing the Economy Through Financial Intermediation," paper presented at Carnegie-Rochester Conference, Pittsburgh, Nov. 14–15, 1975, unpublished.

S. **Fischer,** "The Demand for Index Bonds," *J. Polit. Econ.*, June 1975a, *83*, 509–34.

———, "Non-Indexation in the Capital Markets," paper prepared for the International Economic Association Conference on Inflation, Saltsjöbaden, Aug.-Sept., 1975b, unpublished.

M. **Friedman,** "Using Escalators to Help Fight Inflation," *Fortune*, July 1974, *90*, 90–97+.

D. **Lessard and F. Modigliani,** "Inflation and Residential Financing: Problems and Potential Solutions," in *Innovations in Mortgage Design*, Federal Reserve Bank of Boston Conference Series, *14*, 1975.

W. **Poole,** "Housing Finance Under Inflationary Conditions," in Federal Reserve Staff Study, *Ways to Moderate Fluctuations in Housing Construction*, Board of Governors of the Federal Reserve System, 1972, 355–76.

A. **Rubner,** "The Abdication of the Israeli Pound as a Standard of Measurement for Medium and Long-Term Contracts," *Rev. Econ. Studies*, Oct. 1960, *28*, 69–75.

# ECONOMIC STATUS OF WOMEN IN INTERNATIONAL WOMEN'S YEAR

# The Economic Status of Women in the United States

*By* MARY DUBLIN KEYSERLING*

International Women's Year (*IWY*)—1975— has provided an extraordinary opportunity for stock taking with respect to where women are in the American economy, where they aren't, and where they ought to be. It has helped focus national attention on the many barriers which still impede the full contribution of women; it has intensified commitment to needed action.

Today, some 38 million women, aged 16 and over, are in the American civilian labor force. The number of female workers has more than doubled since 1950 and has nearly tripled since 1940—one of the most far reaching social and economic transformations of our history. As of October 1975, 47 percent of all women, 16 and over, were in the civilian labor force, as compared with 33 percent in 1950 and 29 percent in 1940. (Of all women aged 18–64—a more meaningful base—the proportion of those in the labor force is now 55 percent.)

Between 1950 and 1974, the total labor force grew by 29 million. Women constituted three out of five of this increase. The economy was expanding, although not enough; the number of jobs was growing; and women, particularly mothers who aspired to higher living standards for their families, entered the job market in rapidly rising numbers.

* Economic consultant, former director, Women's Bureau, U.S. Department of Labor.

Prior to World War II, the majority of young women, on leaving school and college, sought jobs. In 1940, women's peak labor force participation rate of 48 percent was reached between the ages of 20 and 24. From the age of 25 onward, the likelihood of women working diminished rapidly, and by the age of 45–54, less than one quarter of those in the age group were wage and salary earners.

Today, the peak labor force participation rate is still reached by those 20–24 years old; 65 percent of women of this age are workers. But the most significant contrast with the prewar period is, that after their mid-twenties, far fewer women drop out of the labor force and, when they do, their absence is much shorter. The civilian labor force participation rate is presently 57 percent for women aged 25–34 as well as for those aged 35–44, and it drops off only one percentage point for those aged 45–54. At age 55–64, 2 out of 5 women are still wage earners, a proportion only a little lower than it was 25 years ago among those most likely to seek employment—the 18–24 year olds.

The increase in the labor force participation rate of mothers of children under 18 has been exceedingly rapid; it quintupled between 1940 and 1974. The availability of jobs, the reduction in family size, and the rising cost of child rearing were major contributing factors.

Women, especially mothers, work for

the same reasons as men. They need the money. Nearly three quarters of all working women are single, widowed, separated or divorced, or have husbands with incomes of under $10,000. Employment is a compelling economic necessity for the 2½ million working mothers of children under 18 who head their own families. The earnings of more than half of the 11 million employed mothers living with their husbands, and who have children under 18, take their families out of poverty or help assure them of a minimum adequate standard of living.

As of March 1975, 55 percent of mothers with children aged 6–17 only, were wage and salary earners—a participation rate that had nearly doubled since 1948. During this period, the rate for mothers with children under the age of six tripled, rising from 13 percent 27 years ago to 39 percent last year.

This development has created an urgent need for action. More than 6½ million children under the age of six have mothers who are gainfully employed. As a nation we have done very little to assure the provision of adequate care for them while their mothers must be away from home. This has become a national responsibility because only a relatively small proportion of families in which mothers of preschool children work can afford the cost of developmental child care. In consequence, a very large proportion of their youngsters are in essentially custodial care, some of it so poor as to be actually injurious. Our record, as a country, with respect to meeting day care needs, is shockingly inadequate compared with most other industrialized nations.

Licensed child care facilities have the capacity to care for only about one million children, and many of these homes and centers fail to meet generally agreed upon minimum standards. Our fish and wildlife departments can tell you the number of fish by species in all our major lakes, but we have no up-to-date official statistics on where most children are or how they are cared for—the last federal study of child care arrangements having related to 1965. Based on that study, undertaken jointly by the Women's and Children's Bureaus, we can estimate that about one-third are cared for in other people's homes and about 10 percent are in group centers. Perhaps as many as one-half the preschool children of working mothers are cared for in their own homes, although few working parents can make satisfactory in-home care arrangements, at a price they can afford. In many homes, fathers who work at night and try to sleep during the day, watch out as best they can for their youngsters—hardly developmental attention. Many studies have revealed that large numbers of school-age children are kept home from school to look after younger siblings. An estimated 10 percent go with their mothers to their places of work. The number of latch-key tiny tots on their own is acutely disturbing.

No less distressing is the almost total lack of before and after school care for young school-age children whose mothers work. What is more, the gap between the supply of adequate child care services and the need for them is widening.

Under the able leadership of Senator Walter F. Mondale, constructive legislation was introduced and was passed by both Houses of Congress only to be vetoed by President Nixon in December 1971. A more modest bill is now pending but its prospects have been diminished by the well-financed opposition of a few, small groups one of which was largely responsible for the language of the Nixon veto. Another is disturbing flyers widely, falsely claiming that the bill would take

children from their parents, make them wards of the State and give them the right to sue their parents.

The plight of children and the anguish of working parents who can't afford to purchase good care can no longer be ignored. And working parents are not the only ones who want and need the expansion of good facilities which requires public support. Part day services should be available especially for children in economically disadvantaged families in which the mothers are not gainfully employed but who want their children to have development opportunities beyond those available at home. Only a few hundred thousand are now in programs such as Head Start. There are very few facilities for the children of students or of nonemployed mothers who, because of health or other problems, cannot care adequately for their children during the day.

## I. Women's Occupational Status

Another urgent problem with respect to women's economic status is the serious underutilization of their skills and talents. We are moving far too slowly to redress it.

Employed women today are heavily concentrated in the lower paid, lesser-skilled occupations which they have traditionally held. In 1973, more than a third of them held clerical jobs and more than a fifth were in service trades with average annual earnings in these two areas, for year-round, full-time work, of about $6,500 and $4,100, respectively. More than an additional fifth were sales, craft and kindred workers, and operatives, with mean average earnings in these three fields combined of less than $5,000 (year-round, full-time).

Fifteen percent of working women are in the professional and technical category, but nearly half of them are concentrated in three traditionally female, and relatively low paid, occupations: elementary and preschool teachers, registered nurses and health technologists.

Women's employment pattern remains very different from that of men. They comprise 40 percent of the total labor force but they are 98 percent of all registered nurses, and also of household workers, 85 percent of all elementary school teachers, 77 percent of those in clerical occupations, and 58 percent of nonhousehold related service employees. And three quarters of all working women are in these five fields, characterized by low average earnings.

As was earlier stated, 60 percent of the increase since 1950 in the total number of workers has been comprised of women. That they have moved even more concentratedly into the relatively disadvantaged jobs is indicated by the fact that their median wage and salary income for year-round, full-time work fell from 64 percent that of men in 1955 to 57 percent in 1974. And the gap has widened in every major occupational group with the one exception of service jobs, excluding private household employment.

In 1973, the median wage and salary income for year-round, full-time work was $6,448 for women and for men—$11,306. Nearly a third of these women earned less than $5,000, as compared with fewer than 1 out of 10 men. At the other end of the income range, only 2 percent of the women earned $15,000 to $25,000; the percentage for men was nearly ten times higher. And at the top of the ladder, less than one third of one percent of the women earned salaries exceeding $25,000; the proportion of men who did so was more than 18 times higher.

Some contend that encouraging employment trends are in process for women because the fastest increase in their

employment in any of the 9 major occupational groups, during the past 20 years, has been in the professional and technical occupations. It is true that women's numbers have more than doubled in these fields during this period, but the number of men in them has increased almost as rapidly. Women represented 38 percent of those in these relatively advantaged occupations in 1955; the proportion subsequently increased by only 3 percentage points—to 41 percent in 1973, but it is still some distance from the 1945 ratio of 47 percent.

It should be noted that in recent years women have become even more crowded in the lower ranks of the professions. In 1961, the median earnings of year-round, full-time professional and technical women workers were 68 percent those of men in such employment. In 1973, the ratio, a measure of relative occupational position, was down to 64 percent.

At the more privileged end of the professional ladder, women remain poorly represented. In 1974, women were 7 percent of all lawyers and judges, 10 percent of physicians, 14 percent of all chemists, 20 percent of computer specialists and 31 percent of all college and university teachers, to mention a few examples.

Even within these underrepresented professions women concentrate at the lower end of the scale and some trends are actually regressive. For instance, in our colleges and universities in 1974; women were 9 percent of all full professors—down from 10 percent in 1959–60; they were 15 percent of associate professors—down from 17.5 percent in 1959–60; and they were 24 percent of assistant professors and 45 percent of all instructors. These last two ratios were higher than in the base year but these gains did not compensate for the declines at the higher levels, so that, overall, women lost salary ground relative to men. At each of the teaching levels the average compensation of women was lower than that of men—the difference for all ranks combined averaging 18 percent.

Another illustration is employment in the Federal Civil Service. While Uncle Sam might well be expected to be the pacesetter with respect to affirmative action policy, being charged by an Executive Order to prohibit discrimination on the basis of sex in federal employment, an analysis of Federal Civil Service reports indicates that the mandate is far from fulfilled. Women in 1973 were only 34 percent of total white-collar federal employees—the same proportion as in 1968, and one much lower than their proportion in the total labor force. Not only had their number in the Service failed to keep pace with their growing role in the general economy, such progress as they have made at the top, in recent years, has been very slight.

In 1973, women were only 2.4 percent of all those in the three highest grades—GS 16, 17 and 18—up from 1.5 percent in 1968. They were 4 percent in the next two highest grades—GS 14 and 15—up from 3.4 percent in 1968; and they were 7.5 percent of those in the middle ranks, GS 12 and 13—up from 5.7 percent five years earlier.

That competent women are available to fill higher jobs needs no argument. Suffice it to say that in the government of the District of Columbia, which I serve as Chairperson of its Commission on the Status of Women, women comprise 10 percent of those in grades 16–18, 15 percent in grades 14 and 15, and 25 percent in grades 12 and 13 (all at the same pay scale as in the Federal Service). This is a record, I might add, many times better than that of the federal government, but it is one with which my Commission is rightfully far from satisfied.

Clearly, discriminatory hiring and pro-

motion practices seriously impede women's employment progress, despite the very encouraging legislative gains achieved since 1963. One of the many challenges is the need for far more effective enforcement of the Equal Pay Act, of Title VII of the Civil Rights Act, of State Fair Employment Practices and Equal Pay Laws, of Executive Order 11246 (as amended by Executive Order 11375) which prohibits discrimination on the basis of sex by federal contractors and subcontractors, and on federally assisted construction projects, of Executive Order 11478 which prohibits discrimination in Federal employment, and of Title IX of the Education Amendments of 1972 which prohibits sex discrimination in federally funded education programs and activities. Enforcement must also be better financed. The backlog of cases is of serious magnitude. More women need to bring their complaints to the agencies as individuals and as groups, and brings suits into court.

Educational efforts are vital to change employer attitudes as well as practices. Parents, teachers and vocational guidance personnel need to be far more aware of their responsibilities to help young women anticipate the realities of their work futures, to set their sights higher, and to prepare to use their abilities to the full. Too many of our girls are steered into traditional deadend jobs. Training programs must be desegregated and greatly improved. Blatant sexism in teaching materials in current use in many schools must be eliminated.

This is not to say "all women should work." Freedom of choice remains basic, to the extent economic considerations make it possible. But the reality is that the majority of women are now jobholders and their proportion will continue to increase. It is a terrible waste for them and for society that brains, skills and energies that have no gender, should be as conspicuously underutilized as they are today.

It is sometimes suggested that because of women's family as well as job responsibilities, greater discontinuity in their employment is inescapable and that this will continue to make for job status sex differentials. But intermittency in women's employment has greatly diminished and its extent is frequently exaggerated. Men are assuming more responsibilities in the home. This trend will accelerate. Few adequate studies have been made, but such data as are available indicate that the gap between the quit rates of men and women, when properly compared, is not wide and has been narrowing. Interestingly, in occupations in which there are large numbers of both men and women, recent studies indicate that occupational mobility rates were higher for men than women.

There are no differences in the overall levels of educational achievement of men and women sufficient to account for the great differences in their employment patterns. There is virtually no gap in the mean number of years of schooling completed—12.1 years by women workers and 12.0 by men. In 1974, a higher percentage of women workers than men had completed high school but a slightly smaller percentage had had some college education—28 as compared with 31 percent. In the area of higher professional occupations where advanced degrees relate to career achievement, it should be noted that women received 43 percent of all masters degrees last year—a somewhat higher proportion than they represented in the labor force, and an all-time peak.

Differences with respect to the percentage of doctoral and first professional degrees received by women do play an important role in upward mobility at some of the highest job levels. In 1974, women received 19.1 percent of all Ph.D.s.

Might I add that in economics, women are not doing as well as we should. In 1974 we received only 9.5 percent of the doctorates granted, but perhaps there is some comfort that this is about double the 1960 proportion.

With respect to first professional degrees there have been some dramatic breakthroughs that bode well for the future relative job status of women in the professions.

During the period 1969–74, the percentage of women among first year enrollees in the medical schools rose from 9 to 22 percent, and in law from 6 to 23 percent. By 1974, women were 31 percent of first-year students in pharmacy and 25 percent in veterinary medicine. There are lags, however, in other professional fields. Women are still only 6 percent of students enrolled in schools of engineering, 7 percent in dentistry, 9 percent in architecture and 10 percent in optometry.

While we are mentioning breakthroughs, a few interesting developments in nonprofessional fields might be referred to. Between 1960 and 1970, the number of women working as craft and kindred workers increased from 277,000 to 495,000—a rate of increase 8 times that for men in these skilled trades. They include carpentry, for instance, in which the number of women rose from 3,300 to 11,000; women electricians increased from 2,500 to 8,700; plumbers from 1,000 to 4,000; auto mechanics from 2,300 to 11,000; and tool and die makers from 1,100 to 4,200. But lest too many rise up to cheer, let me caution that women remained only 1 to 2 percent of all those employed in these five trades.

There is another heartening note. While, as I stated previously, the earnings gap between men and women has been widening, the gap between the earnings of black and white workers have become considerably smaller over recent years.

In 1939, the median year-round, full-time earnings of nonwhite women was 38 percent those of white women. By 1973, the ratio had risen very sharply to 88 percent, reflecting diminishing racial differences in the occupational patterns of the two groups. Measured in constant dollars, the earnings of nonwhite women, for year-round, full-time work, increased 5½ fold from 1939 to 1973. They needed to; they were so tragically far behind. Minority men's earnings, which have been 45 percent those of white men, rose by 1973 to 72 percent.

## II. Unemployment is Higher for Women than Men

Generally, the lower the status of the job, the more subject it is to unemployment. This is one of the main factors accounting for the greater vulnerability of women to unemployment, which averaged 30 percent higher for them than men during the past 25 years. Another element is seniority practices leading to the first firing of the last hired, which affect women disproportionately, especially minority women. These rules should be adjusted when they perpetuate past discrimination or impede affirmative action goals.

Officially recorded unemployment among women reached a post-World War II peak at the high rate of 10.2 percent in May 1975. As of October, the rate was still 9.1 percent. (Both May and October rates are seasonally adjusted.) The comparable rates for men were 8.5 and 8.2 percent, respectively. Unemployment among women during the first 10 months of 1975, averaged 57 percent higher than in 1973.

## III. Some of the Special Economic Problems of Minority Women

Before coming to a close, I want to add a few additional comments about the relative economic status of white and minority

women. Of all women in the labor force in 1974, 32.9 million were white and 4.9 milion, or 13 percent, were members of minority races (of which about 90 percent were black).

For many years the labor force participation rate had been higher for minority than white women. In 1947, it was 47 percent higher. That year 46.9 percent of minority women were in the civilian labor force, as compared with 31.8 percent of white women. Subsequently, this difference markedly diminished. By 1974, the labor force participation rate of nonwhite women was 49.1 percent, but that of white women has risen far more rapidly and was 45.2 percent. The differences in these growth rates would, undoubtedly have been less had minority women not been far harder hit by unemployment.

Recorded unemployment among minority women was 78 percent higher than among white women during the period 1949–74, averaging 9.1 percent as compared with 5.1 percent. And, it should be stressed, official unemployment rates are more understated for minority women.

The larger financial responsibilities of minority women is another factor which might have lifted their employment rates faster, had more jobs been available. The average earnings of nonwhite males are lower than those of whites, as earlier discussed, which puts added pressure on minority wives to seek work. Further, a far larger and more rapidly growing proportion of minority women than white, head their families. In 1974, more than a third of all minority families were headed by women, as compared with one out of ten white families. (The respective proportions had been 24 and 9 percent in 1965.)

Despite these comparative difficulties, the occupational pattern of minority women has been improving relative to that of white women. In 1965, 55 percent of all employed minority women were in service occupations; this proportion had dropped sharply to 37 percent by October 1975. Today, a much larger proportion of women in the relatively more advantaged white-collar jobs are members of minority races—their proportion having risen from 25 to 43 percent during the past ten years.

But there are still marked occupational differences between the two groups of women. While 37 percent of minority women are in service occupations, this is nearly twice the proportion of white women in this occupational category. And while 43 percent of minority women are now in white-collar jobs, this compares with more than 65 percent of white women. These, among other differences indicate the continuing need for the intensive acceleration of efforts to eliminate racial disadvantages.

That occupational differences between the two groups of women may be expected to narrow to a greater degree in the years immediately ahead is indicated by the fact that the gap between the occupational status of white and minority women, aged 16–34, is very much narrower than between women 35 years of age and older. This reflects progress especially with respect to relative educational opportunities. There is now virtually no difference between the median years of school completed by white and minority women workers (12.5 and 12.3 years, respectively, in 1974). Only 15 years ago, there was a difference of 2.8 years. There is much still to be done, nonetheless to assure the equalization of education with respect to quality.

This brief comparison of white-nonwhite economic status cannot be concluded without some mention of the special unemployment problems of minority teenagers, both male and female. Their recorded unemployment rate is unconscionable. It was 32.9 percent last year

and to the outrageously high postwar peak of 41.6 percent in March 1975, when the rate was more than double that of white teenagers. During the past 25 years, to be teenaged, black and female has added to the unemployment hazard; their unemployment rate has averaged 25 percent higher than for nonwhite boys.

Including the estimated number no longer actively searching for jobs, and hence not counted as unemployed, nearly one out of two nonwhite teenage girls who want work are now jobless. This is a miserable way to enter adult life in this the richest country on earth, fully capable of providing jobs for all.

## IV. Additional Targets for Action

Every year from next year forward should be International Women's Year until today's glaring inequities are eliminated. I have already touched on some much needed targets for action—including the need for increases in services, more effective administration and enforcement of laws and orders prohibiting sex and racial discrimination, and educational and other efforts to change attitudes and practices which now perpetuate employment barriers. Let me mention just a few additional challenges.

There is urgent need for reform of our social security system to eliminate inequities. Women are more likely to collect higher benefits as their husbands' dependents than as workers in their own right.

Our unemployment insurance system also needs reform. In more than 30 states, women are disqualified from compensation because of pregnancy, regardless of their ability to work. Part-time workers are not covered by unemployment compensation—and 26 percent of all women work part time—a third of them year round. Some 10 million workers, mainly women are still excluded from state unemployment insurance programs.

Many inequities with respect to taxation should be eliminated. The cost to working women of child care services should be deductible as a business expense.

Minimum wage rates lag behind soaring living costs. They should be lifted to make life more liveable for many of our most deprived families, the majority of which are headed by women. And it is imperative that an income maintenance program that would assure all families at least an adequate minimum living standard, replace our welfare system.

The Equal Rights Amendment must be ratified.

Women in far larger numbers should run for political office at the national, state and local levels. There is a long way to go when so few are in Congress, and when only 8 percent of our state legislators and 4 percent of the mayors of our 1000 largest cities are women. Women must press for far more adequate representation on appointive boards and commissions.

And there is one central underlying goal on which virtually all others depend. We must recognize that a healthy economy, growing at the rate it is capable of, is essential if there are to be jobs enough, opportunities for economic advance, and economic justice for all our people. The price we pay for the waste of our human resources when economic stagnation and recessions recur, is incalculable.

All these goals call for more than words. They call for mobilization for action for which economists have a very special responsibility.

1976 is Bicentennial Year. If we accept our responsibility, let us hope it will not be long before we shall be able to say that the Declaration of Independence has at long last been extended to all our citizens—holding the truth to be self evident that all men—and all women—are created equal.

# The Economic Position of Women in the United Kingdom

*By* NANCY SEEAR*

For women, as for men, it is the ownership of wealth and the acquisition and maintenance of income which together determine their economic position. This paper attempts to establish, so far as available information permits, the nature of the existing economic situation of women in the United Kingdom at the present time, to explore the reasons for the position, and to examine, tentatively, factors which may lead to changes over the next decade.

Accurate estimates of the distribution of wealth are notoriously difficult to calculate, but according to the recently published report of the Royal Commission on the Distribution of Income and Wealth, in 1927 women in England and Wales owned 33.1 percent of total wealth, the figure rising to 42.2 percent in 1962 but falling again to 39.4 percent in 1973. In the year 1973 there were 20,400,000 males aged 15 and above and 22,800,000 females aged 15 and over in the United Kingdom (i.e., Great Britain and Northern Ireland). In the same year it was calculated by the Inland Revenue that total personal wealth in Great Britain amounted to £163,900 million, distributed among 19.1 million wealth holders. Of these wealth holders 11,009,000 were men and 8,131,000 were women, and of these 4,087,000 or approximately 63 percent of the men, as compared with 4,316,000 or 52 percent of the women, owned under £3,000, while 2,759,000 or 25 percent of the men but only 1,617,000 of the women owned £10,000 or over.

* London School of Economics.

From these figures it is apparent that for the great majority of women it is on income rather than on ownership of wealth that they must rely almost entirely to maintain their economic position. To a lesser extent this is also true of men. It is estimated that in 1973 the bottom 80 percent of the total population, male and female, aged 18 and over, owned 13.6 percent of the total personal wealth in the country.

These figures must be interpreted cautiously. Different methods of calculation and the exclusion of certain categories of wealth from the Inland Revenue statistics on which they are based can lead to considerable variations in the estimates. But however calculated, it is clear that the ownership of wealth is of marginal importance except for the few. In assessing these figures it must on the other hand be borne in mind that 52 percent of all householders owned or were purchasing their own house in 1973. This is of particular value to married women, whether or not the house is in joint ownership. Women's expectation of life is longer than that of men. There are more than double the number of women than men over the age of 80. An increasing number of married women can therefore look forward to the ownership of their own home during their period of widowhood.

Clearly for the great majority of women it is the ability to earn a salary or wage and to establish rights under public or private social security schemes which is of the greatest importance. How do women fare in these respects?

TABLE 1—GREAT BRITAIN

| | Median Earnings | Lowest Decile | Lower Quartile | Median | Upper Quartile | Highest Decile |
|---|---|---|---|---|---|---|
| | £ per annum | % | % | % | % | % |
| Men | | | | | | |
| 1964/65 | 934 | 65.2 | 80.8 | 100.0 | 123.6 | 155.9 |
| 1968/69 | 1,198 | 64.3 | 80.2 | 100.0 | 125.0 | 159.1 |
| 1972/73 | 1,861 | 64.0 | 79.6 | 100.0 | 124.9 | 158.8 |
| Women | | | | | | |
| 1964/65 | 466 | 55.7 | 76.5 | 100.0 | 131.6 | 177.8 |
| 1968/69 | 583 | 52.9 | 74.5 | 100.0 | 132.7 | 179.0 |
| 1972/73 | 907 | 53.4 | 73.1 | 100.0 | 134.4 | 178.3 |
| Men and women | | | | | | |
| 1964/65 | 825 | 47.7 | 69.9 | 100.0 | 130.0 | 165.1 |
| 1968/69 | 1,034 | 45.3 | 68.0 | 100.0 | 132.3 | 168.8 |
| 1972/73 | 1,582 | 44.7 | 67.5 | 100.0 | 133.6 | 170.4 |

*Source:* Royal Commission on the Distribution of Income and Wealth, report no. 1. Derived from DHSS Statistics of Earnings.

Table 1 shows the distribution of earnings for both men and women in the years 1964–73. They cover the annual earnings of men and women aged 18 or over in full- or part-time employment throughout the year, median earnings, and deciles and quartiles as percentages of the median for 1964–65 to 1972–73. Median earnings distinguishing between manual and nonmanual workers, and between men and women for the years 1970–74 are given in Table 2.

These figures are not strictly comparable with the figures in Table 1 as they are based on the gross weekly earnings of full-time males aged 21 and over and full-time female employees aged 18 and over, whose pay for the survey period was not affected by absence.

In comparing these earnings figures, it has to be remembered that earnings are increased by overtime and shift working which is undertaken by men to a much greater extent than by women. In 1971,

TABLE 2—MEDIAN EARNINGS
(£ per week)

| | Manual Men | Nonmanual Men | All Men | Manual Women | Nonmanual Women | All Women |
|---|---|---|---|---|---|---|
| 1970 | 25.6 | 31.4 | 27.2 | 12.8 | 15.9 | 14.6 |
| 1971 | 28.1 | 34.4 | 29.8 | 14.6 | 18.0 | 16.6 |
| 1972 | 31.3 | 38.5 | 33.4 | 16.4 | 20.1 | 18.6 |
| 1973 | 36.6 | 42.8 | 38.4 | 18.9 | 22.3 | 20.9 |
| 1974 | 41.8 | 48.5 | 43.8 | 22.7 | 26.1 | 24.7 |

*Source:* Royal Commission on the Distribution of Income and Wealth, report no. 1. Derived from NES, 1974, Table 15.

for example, so called normal working hours for manual workers were 40 for both men and women. But in the industries and sources covered by the regular earnings enquiries, which include all manufacturing industries, men averaged 44.7 hours and women 37.7. If women officially classified as part-time workers, i.e., working 30 hours or less are included, as they are in Table 1, the actual average working hours for women dropped to 32.4.

In 1971, only 3.9 percent of all male workers, but 33.7 percent of all female workers, worked 30 hours a week or less, while 47.8 percent of men but only 12.6 percent of women worked over 40 hours. Even if women officially classified as part-time workers are excluded, as in Table 2, only 22.4 percent of women manual workers worked 40 hours or more compared with 64.5 percent of male manual workers. The contrast among nonmanual workers is far less sharp. Among nonmanual full-time workers, men averaged only 1.7 hours more than women and about 40 percent of both sexes worked 36–38 hours. Even among nonmanual full-time workers however, 30.9 percent of women but only 19.2 percent of men worked 36 hours or less, while 8.6 percent of men but only 1.6 percent of women worked 45 hours or more.

With these reservations, the Department of Employment, in summarizing trends was able to comment:

> In recent years average weekly and hourly earnings of women have increased relatively faster than those of men, both among non-manual and among manual employees . . . The New Earnings Survey shows, for example, that between April '72 and April '73 the percentage increases in average weekly earnings, excluding overtime earnings for full-time manual workers were 16.2% for women aged 18 and over but 15.1% for men aged 21 and over; and for full-time non-manual workers they were 13.7% for women but 12.8% for men. [p. 30]

Women's economic position, insofar as it depends on earnings is therefore relatively, if slowly improving. But with a median earnings figure of £24.7 in 1974 it can hardly be regarded as strong.

If women's income is not high, can it be considered secure? Security depends mainly on two factors: on the relative reliability of employment, and on the provision available to plug the gap when earnings are interrupted or cease.

In the United Kingdom in recent months, unemployment figures have been rising to a level not experienced since World War II. Contrary, however, to the expectations of some commentators the employment of women does not appear to be affected worse than that of men. This is perhaps not surprising since women are to so large an extent confined to customary women's jobs. They are therefore complementary to and not in competition with men in the labour market. If and when the aims of the Sex Discrimination Act are achieved, and jobs are open indiscriminately to men and to women, the protection afforded by the present organization of the labour market may well disappear, and women may in practice be more vulnerable, despite the safeguards in the Equal Pay Act and in the Sex Discrimination Act.

Though women are not more likely to experience unemployment than men, their position is much less favourable when, for any reason, earnings disappear. This is in part due to a curious provision in British Social Security legislation which has permitted married women employees to contract out of payment of Social Security contributions, except in relation to industrial injuries. In 1971, 61 percent of all women employees were married and about 75 percent of all married women employees took advantage of this right to exemption. A very high percentage of women employees are therefore not eligible for

unemployment benefit, for sickness benefit or for a retirement pension in their own right. This option has been removed in the Social Security (Pension) Act 1975, but not for married women already in employment.

As a result the great majority of married women are dependent, not only in sickness and unemployment, but also in old age on benefits earned by virtue of their husbands' contributions. These contributions provide the husband with a dependent's allowance for his wife, at a level considerably less than the pension the wife could have obtained had she taken up the option to be insured in her own right. She is also entitled to a widow's pension on the strength of her husband's contribution. If however, she is divorced and the husband remarries, it is the current wife, not the original wife, who qualifies for the widow's pension, though the first wife can make use of his contributions up to the time of the divorce to satisfy the contribution requirements for her own pension. This unsatisfactory situation underlines the importance of introducing social security policies which establish women's social security rights on an individual basis, regardless of marital status. The maintenance of income when earnings are interrupted does not of course depend entirely on payments from the National Insurance or Social Security funds. Increasing fringe benefits including especially company sick-pay schemes and occupational pension schemes are contributing an increasing proportion of income. How then do women fare in these nonstatutory schemes?

The Equal Pay Act 1970 required that there should be no discrimination by employers on grounds of sex in relation to fringe benefits such as sick pay and holiday entitlement, but explicitly excluded pensions and superannuation rights, the area which is of far the greatest importance. Under the Social Security (Pensions) Act of 1975, it is a condition of approval of a company superannuation scheme by the statutory Occupational Pensions Board, that there should be no

TABLE 3—PERCENTAGE OF FEMALE WORKERS IN MAJOR OCCUPATIONAL GROUPS, 1911–66

| Occupational Groups | 1911 | 1921 | 1931 | 1951 | 1961 | 1966 |
|---|---|---|---|---|---|---|
| 1. Employers and proprietors | 18.8 | 20.5 | 19.8 | 20.0 | 20.4 | 23.7 |
| 2. White-collar workers | 29.8 | 37.6 | 35.8 | 42.3 | 44.5 | 46.5 |
| (a) Managers and administrators | 19.8 | 17.0 | 13.0 | 15.2 | 15.5 | 16.7 |
| (b) Higher professionals | 6.0 | 5.1 | 7.5 | 8.3 | 9.7 | 9.4 |
| (c) Lower professionals and technicians | 62.9 | 59.4 | 58.8 | 53.5 | 50.8 | 52.1 |
| (d) Foremen and inspectors | 4.2 | 6.5 | 8.7 | 13.4 | 10.3 | 11.4 |
| (e) Clerks | 21.4 | 44.6 | 46.0 | 60.2 | 65.2 | 69.3 |
| (f) Salesmen and shop assistants | 35.2 | 43.6 | 37.2 | 51.6 | 54.9 | 58.7 |
| 3. All manual workers | 30.5 | 27.9 | 28.8 | 26.1 | 26.0 | 29.0 |
| (a) Skilled | 24.0 | 21.0 | 21.3 | 15.7 | 13.8 | 14.7 |
| (b) Semi-skilled | 40.4 | 40.3 | 42.9 | 38.1 | 39.3 | 42.6 |
| (c) Unskilled | 15.5 | 16.8 | 15.0 | 20.3 | 22.4 | 27.5 |
| 4. Total occupied population | 29.6 | 29.5 | 29.8 | 30.8 | 32.4 | 35.6 |

*Source:* G. S. Bain and R. Price, "Union Growth and Employment Trends in the United Kingdom 1964–1970," *Brit. J. Ind. Rel.*, Nov. 1972; reproduced with the permission of the editor.

discrimination on grounds of sex in such schemes and this is reflected in the state superannuation provisions. This will ultimately ensure income-related superannuation payments to all persons, male and female, who have been in paid employment. The position of widows is also to some extent safeguarded by the requirements of the Occupational Pensions Board.

Since earnings and the fringe benefits dependent on earnings are decisive, the opportunities available to women in the labour market are crucial. In the United Kingdom, as to a greater or lesser degree in all western democracies, there are virtually two labour markets, a men's labour market and a women's labour market. The distribution of jobs has been dominated by the assumption that there are men's jobs and women's jobs with little if any scope for interchangeability. There are two major characteristics of the women's labour market: women are concentrated in a small number of occupations and are to be found to a marked degree in the lower levels of employment.

> Analysis by occupational order of the 8,344,100 female employees in employment showed that 2,429,800 (or 29.1%) of all female employees were 'clerical workers'; 1,931,700 or 23.2% were service, sports and recreation workers (e.g. canteen assistants, cooks, office cleaners); 994,600 (11.9%) were professional, technical workers and artists (e.g. teachers, nurses); and 896,400 (10.7%) were 'sales workers.' These low occupational orders in total accounted for nearly three quarters of all female employees.[1]

Table 3 shows the underrepresentation of women in the more responsible, interesting and skilled jobs, while Table 4, giving the number of girls entering the apprenticeships to skilled occupations in Great Britain in 1972 makes it clear that no great change can be expected in the near future in view of the small number of girls entering training for skilled work on leaving school.

TABLE 4—GIRLS AGED 15–17 ENTERING APPRENTICESHIPS TO SKILLED OCCUPATIONS IN 1972—GREAT BRITAIN

| | Number | Girls as Percentage of Total in Each Trade Group |
|---|---|---|
| All manufacturing industries | 814 | 2.5 |
| Distributive industry | 804 | 12.1 |
| Insurance, banking, finance and business services | 68 | 14.9 |
| Professional and scientific services | 1,137 | 42.6 |
| Miscellaneous services of which: | 14,654 | 47.8 |
| Hairdressing and manicure | (13,969) | (94.0) |
| Other industries | 520 | 1.2 |
| Total | 17,997 | 15.2 |

*Source:* "Returns from Careers Offices," *Department of Employment Gazette*, May 1973.

Table 5 demonstrates how far women lag behind in the professional field.

This is a familiar picture. The reasons for it are in the main similar on both sides of the Atlantic. Deep rooted cultural assumptions are no doubt at the heart of the matter, assumptions which are reflected in influences in the home, in the behaviour encouraged in girls in contrast to boys, in educational provision and in employers' and trade unionists attitudes. These attitudes are so deep seated that they are frequently totally unrecognized: unlike racial prejudices, their recognition, when it occurs, is rarely accompanied by any feeling of guilt. It is also true that such traditional attitudes are often found in women as well as in men. In a study which I carried out in 1962, women in positions of responsibility were asked about their attitude toward women bosses. Those who had never experienced a woman boss overwhelmingly held the traditional view that a male boss was to be preferred. Inter-

[1] This extract is based on the 1 percent sample results of the 1971 Census of Population.

TABLE 5—WOMEN IN CERTAIN PROFESSIONS IN 1971 AND 1972[a]

| Institute | Total Membership (number) | Percent Who were Women |
|---|---|---|
| Institute of Chartered Secretaries and Administrators | 43,954 | 2.6 |
| Chartered Insurance Institute | 50,079 | 3.7 |
| Institute of Bankers | 30,128 | 1.2 |
| Institute of Chartered Accountants in England and Wales | 52,738 | 1.6 |
| Institute of Chartered Shipbrokers | 3,486 | 0.4 |
| Institute of Electrical Engineers | 30,710 | 0.1 |
| Institute of Mechanical Engineers | 73,851 | 0.1 |
| Law Society | 25,366 | 3.2 |
| Royal Institute of British Architects | 20,394 | 4.2 |
| Royal Institution of Chartered Surveyors | 31,203 | 0.4 |
| Royal Town Planning Institute | 4,519 | 5.4 |
| British Medical Association | 49,714 | 17.8 |
| Institute of Civil Engineers | 19,424 | 0.0[b] |
| Royal Institute of Chemistry | 26,172 | 3.8 |
| Institute of Structural Engineers | 13,804 | 0.1 |
| Institute of Building | 20,964 | 0.0[c] |
| Institute of Heating and Ventilating Engineers | 5,965 | 0.3 |
| British Dental Association | 12,200 | 13.1 |
| General council of the bar of England and Wales | 2,877 | 6.4 |

[a] The figures were provided at various dates between April 1971 and September 1972. Membership categories vary but are the same for both men and women in each association's estimates.

[b] There were 7 women members.

[c] About 5 noncorporate members were women.

estingly, among those women who had actual experience of a woman boss, the predominant attitude changed, not to a preference for a woman, but to indifference to the sex of the boss. At the present time when additional resources are being made available under the British Manpower Services Commission for the retraining of women returning to the labour market, a very high percentage of women taking advantage of the new grants are choosing to train for traditional women's jobs.

Attitudes are of prime importance. But at a different level, lack of knowledge also contributes substantially to the present position. In particular many people, especially many employers, do not appear to appreciate the significance of the important demographic changes of recent decades. At the turn of the century when the vast majority of women worked for a very few years between school and marriage and did not contemplate returning to paid employment, there was a rational case against the investment of time and money in training girls. Today, with a falling birth rate and childbearing normally completed before the age of thirty, that argument has lost a great deal of its force. Because employers have not grasped the full implications of this demographic and social revolution, they have not analyzed the ways in which the new type of female labour force can best be deployed; nor have they realized the economic benefits that can be reaped from the new situation. For example, throughout most of the 1950's and 1960's employers suffered from an acute shortage of draughtsmen. Scarcity pushed up draughtsmen's rates, and perhaps even more damaging from the employer's point of view, the inability to get designs off the drawing board delayed production and postponed delivery dates. During this time women in Sweden were being trained as draughtsmen in such numbers that by the end of the 1960's more than half the Swedish draughtsmen were in fact women. In the United Kingdom on the other hand, only 1 percent of the total number of draughtsmen were women by that date.

In much the same way the full significance of the greatly increased tendency for women to return to work after a short period at home for childbearing has not been appreciated. Women reentering the labour market in their middle thirties have a quarter of a century of working life ahead of them. The investment of time and money in training can clearly be justified in terms of a reasonable likely return. Extensive training will not be

appropriate for, or wanted by, all women. Careful selection for women suitable for such training is needed as it is for men. But there has been little attempt to tackle the task of making better use of this newly available womanpower, except in a piecemeal manner.

Many women, as is often pointed out, want only part-time work. But even these cases, handled with discrimination, can be used far more advantageously than at present. Despite increasing efforts to improve job satisfaction, repetitive or limited jobs remain. Such work is more acceptable on a part-time than on a full-time basis. It is also wasteful both from the employer's and the women's point of view to restrict part-time employment to work of this nature. Though it can be argued convincingly that there are jobs, particularly executive positions in line management which require full-time attendance, there are many others, for example in research and in many types of professional and office work, which in reality need not be full-time. Employers stand to gain from such changes, as able people at present underutilized or not employed at all, can be drawn into employment on these terms. A growing number of employers may also find, as employers in retail distribution found years ago, that there are many situations in which work-loading is such that it is far more economical to employ staff part-time rather than full-time.

What is required is in fact a straightforward management job of assessing requirements and resources and finding ways of making the most effective use of the new supply of womanpower. This, as so often, demands a surrender of preconceived ideas, and the use of the right kind of discrimination. In particular, it requires the abandonment of familiar stereotypes. The advancement of women has suffered severely from generalizations about women's labour turnover rates, women's absence rates, women's attitudes to work, regardless of the long list of other important variables which must be taken into account in attempting to predict the behaviour of either men or women. Once this is done, there is plainly a pay-off both for employers and for the community as a whole. The United Kingdom is in no position to waste its resources, most especially its human resources and the underutilization of British womanpower is a glaring example of such waste.

What then is being done, and what results may be expected from these changes?

Three Acts of Parliament have been passed which provide the foundation for much needed change. The Equal Pay Act 1970 set out "to prevent discrimination, as regards terms and conditions of employment between men and women." A three-pronged attack on unequal pay is made in the Act. If a collective agreement contains a discriminatory clause, the agreement can be altered by the Central Arbitration Committee. If job evaluation exists in an enterprise, it must be applied uniformly to men's jobs and to women's jobs. If there is no scheme of job evaluation, a woman can claim the same pay as a man if, but only if, "her work and theirs is of the same or a broadly similar nature." Whether or not work is of a "broadly similar" depends on the nature, the extent and the frequency of the differences.

This Act, though passed in 1970, only came fully into force on December 29, 1975. As soon as it was passed, men and women concerned with the position of women in the economy began to campaign for equal opportunity legislation to complement the Equal Pay Act. Without such additional legislation it seemed likely that equal pay legislation, welcome though it was in many ways, might in fact adversely affect the position of a considerable

number of women. The merging of the men's labour market and the women's labour market would become a necessity if women, displaced from women's jobs by the operation of equal pay legislation, were to obtain appropriate and satisfactory new openings. The need for this change was underlined by manpower forecasts such as that produced by Colin Leicester then of the Department of Applied Economics at Cambridge. Leicester estimated that between the years of 1960 and 2000 the percentage of the working population in clerical and manual work would fall from 12 percent and 70 percent, respectively to 9 percent and 61 percent while the percentage in professional and technical work would rise from 8 percent to 20 percent. Since women were overrepresented in clerical and lower level manual work, and underrepresented in technical work, it was essential to challenge the sex categorization of these jobs if women no longer needed in the clerical and manual fields were to become available for the growing range of technical jobs.

The campaign for equal opportunity culminated in the Sex Discrimination Act 1975 which also comes into force on December 29, 1975. This Act prohibits discrimination on grounds of sex or marriage in many cases including education, training or employment, with a small number of exceptions based mainly on a fairly tight interpretation of genuine occupational qualifications. An important subsection lays down that if the employer applies a condition or requirement which is such that a considerably smaller proportion of men than of women can comply with it, the condition can be challenged, and the employer has to show that it is "justifiable." This section could be very important and very explosive. What about time-served apprenticeship requirements? What about seniority rules?

These two Acts have been supplemented by the Employment Protection Act 1975, which includes a section on maternity benefit. For the first time in the United Kingdom a woman who fulfills the necessary conditions of service cannot be dismissed for pregnancy, can claim six weeks pay as pregnancy leave and has reinstatement rights for a period of twenty-nine weeks. Although child care provisions are quite insufficient to enable a large number of mothers to take advantage of these new rights, it will make it easier for some at any rate to combine a career and motherhood.

Such is the legislation. How much difference will it make to the position of women, and what changes can be expected in women's economic position, in particular in the employment field?

So far as individuals are concerned, the enforcement of equal pay and equal opportunity legislation is mainly in the hands of Industrial Tribunals, which already deal with such matters as claims under the Redundancy Payments Act, and with which people in employment and their union representatives are very familiar. In addition, an Equal Opportunities Commission has been set up with a broad mandate to work towards the elimination of discrimination, to promote equality of opportunity and to keep the working of the legislation under review. The Commission has powers within certain limits to carry out formal investigations and, where they find discrimination, to serve a nondiscrimination notice to be followed up by investigation to establish whether the requirements of the notice are being carried out. If the Commission is not satisfied, the case can be referred to a county court which may grant an injunction. The Commission can also undertake research and educational work and is already taking steps to see that rights and obligations under the Acts are widely known. The legislation does not

give the Commission or anyone else power to require positive discrimination in favor of women except in the very important field of training.

To an American audience it may seem that these powers are inadequate. There are however reasons why, at any rate in the British context, it seems right to place the emphasis on education, training and persuasion and while providing legal sanctions to keep these sanctions in the background. The existence of the law puts the seal of society's approval on the elimination of sex discrimination. It also undoubtedly draws the attention of educationalists, employers and trade unions to the need to examine their existing practices and to begin a process of change. While this is not yet widespread, there are certainly signs that it is beginning to happen. Above all, as has often been said, legislation can help to change attitudes by changing behaviour, but it is the change in attitudes that is important. In other fields of social policy, it is accepted that individuals do not change attitudes as the result of authoritarian action taken by powerful individuals or organizations. Such action more often breeds evasion or at most a minimum conformity with the law, accompanied by a growing resentment which blocks genuine change. It may be slower to help and encourage those who have to change to work out their own methods and to take upon themselves the task of studying what needs to be done and learning, through experience, the way to do it. It may be slower, but there is always the fable of the tortoise and the hare.

Where women will be in ten years time is hard to predict. It depends in part on the future of the British economy, though it is possible that women, uncommitted to traditional working methods might have a specially useful contribution to make to industrial redevelopment. It depends on the realization by employees and unionists that the removal of discrimination though undoubtedly altering long established customs, can if handled right bring considerable benefits and not only to women. Above all, it depends on women themselves. The first stage must be a great increase in training opportunities. It will be for women to make sure that these opportunities are taken.

## REFERENCES

**G. S. Bain and R. Price,** "Union Growth and Employment Trends in the United Kingdom 1964-1970," *Brit. J. Ind. Rel.,* Nov. 1972.

**Brock, Roberts, and B. N. Seear,** *A Career for Women in Industry,* London.

**Department of Employment Gazette,** "Returns from Careers Offices," May 1973.

**Department of Employment Manpower,** "Women and Work, A Statistical Survey," pap. no. 5, 1, 27, 30, 59.

———, "Women and Work—A Statistical Survey," pap. no. 9.

**Royal Commission on the Distribution of Income and Wealth,** report no. 1, cmnd. 7171, 79, 102, 106.

# Three Years of Self-Paced Teaching in Introductory Economics at Harvard

*By* ELIZABETH ALLISON*

For the past three years Harvard has been experimenting with self-paced instruction (*SPI*) in several sections of its introductory economics course, Economics 10. This paper is a report of the "what, how and for whom" of that experiment: what the benefits (and costs) of *SPI* have been, how *SPI* changes the learning process, and for whom among our students it has been particularly helpful.

The reader will note the lack of the evangelicalism often found in descriptions of educational innovations. This is not intended to suggest that self-paced instruction is without merit; indeed our evidence suggests quite strongly that under the right circumstances, the marginal product of *SPI* is a 10–20 percent increase in scores and that it inspires students to take more courses in economics. However, it is also true that producing the "right circumstances" is initially rather costly in time and energy; that the distribution of the benefits is not uniform; and that after an initial period of enthusiasm, students are not happier in self-paced courses than in conventional courses. In short situations in which an *SPI* system would not dominate a conventional course are plausible.

* Dept. of Economics, Harvard University. For a more extensive description of the self-paced experiment at Harvard, see "Self-Paced Instruction in the Introductory Economics Course," Elisabeth Allison, Harvard Institute of Economic Research, pap. no. 368.

## I. Some Background on the Harvard Experiment

The Harvard experiment with *SPI* began in the fall of 1972. Like most educational experiments, its birth was something of an accident; it was born of dissatisfaction with the mastery students displayed at the end of our introductory economics course and the discovery of an enormously intuitively appealing new educational technique which had worked well in physics and psychology. Briefly, a self-paced course has two central features.

The first is a list of the course objectives, specified in operational terms (these objectives can be very detailed), and the second is a set of exams (often called "unit tests") which the student is invited to try whenever he believes he has mastered an objective. The exam is graded immediately and the student who does not pass the exam is encouraged to return and take another exam on the same material. Full credit is given for passing an exam regardless of the number of unsuccessful attempts and a student's grade is to some extent based on the number of exams passed. Thus it supplies a student with prompt nonthreatening feedback.

The distinguishing features of the self-paced course developed at Harvard were three. First, it seemed appropriate to ask our students to synthesize some of the material themselves; thus we had a com-

paratively few broad units (8 per semester) and fairly high-level objectives. Second, whereas in many self-paced systems, objectives and unit tests replace formal meetings, in Economics 10, sections continued to meet although attendance was voluntary. Third, only upperclassmen or graduate students were used as graders, rather than allowing students who had passed a unit test to serve as graders for that unit test.

## II. A Framework for Evaluation

From its inception we had intended *SPI* as an experiment. Thus we had taken advantage of the decentralized structure of the course (it is taught primarily in sections, but has a course-wide reading list and final exam); *SPI* was introduced into 3 sections of 25–30 randomly selected students each semester over a three-year period. We had instituted, simultaneously, a very extensive, course-wide questionnaire and data gathering program, including admissions profiles and student effort reports.

Originally we had intended to follow the conventional evaluation format, of estimating an achievement equation in which a test score was the dependent variable and *SAT* scores, age, sex, and grade point average (*GPA*), the controls, perhaps improving upon it by collection of better ability measures.

However, the closer we looked the less satisfactory this format looked. Intellectually, the *ad hoc* specification of most achievement equations was unsatisfying. The lack of attention to the simultaneity problem seemed likely to produce serious bias in estimates of the coefficients. Moreover, it was not heuristic, as a good theory almost invariably is. Thus we gradually turned to an alternative framework: production theory. Deciding to evaluate *SPI* by treating it as a new technology in a production process brought several advantages. It opened up a well developed technology, both conceptual and econometric. It provided perspective on the limitations of the conventional approach; the economic education literature largely recapitulates the errors of early production theory. Finally, and most important, it made it clear that the development of a full scale model of the educational process was necessary for any thorough evaluation of an innovation. This is not to suggest that the development of an appropriate model within which to evaluate *SPI* has been accomplished. The lack of a generally accepted theory of learning and a severe errors-in-variables problem made the task of finding an appropriate structure and defining operational counterparts to the theoretical terms most difficult and our specifications are still tentative. Nevertheless, even the preliminary models have yielded some interesting results on *SPI* and on the learning process in general.

The simplest of these models, and the one within which our preliminary evaluation of *SPI* took place began as follows.[1] We assumed that the student (our decision-making unit) allocates his efforts among courses and other activities in such a way as to maximize some function:

$$U = f(Grades_{Ec\ 10}, Grades_{Other\ courses}, Time_{Ec\ 10}, Time_{Other\ courses}, Time_{Other\ activities})$$

This assumption allowed us to derive a 3-equation model which here, in the interests of space, we will merely state. The first line is an effort equation:

$$Effort^{i}_{Ec\ 10} = f(Tastes,\ Expected\ grade,\ Pedagogy).$$

Since this exposition is intended to be merely suggestive, we henceforth make the

[1] The construction and estimation of the Economics 10 production function are described in more detail in "An Educational Production Function for an Introductory Economics Course," Elisabeth Allison, Harvard Institute of Economic Research, pap. no. 563.

unnecessary but convenient assumptions of linearity and additivity and can therefore write the equation for the effort put out by the $i$th student as:

$$(1)\quad \mathit{Effort}^i = a + b_1^i\,\mathit{tastes} + b_2^i\,\mathit{expected\ grade} + b_3^i\,\mathit{teacher\ quality} + b_4^i\,\mathit{SPI}$$

The $i$th student chooses to devote some effort, $E^i$, to Economics 10. He faces a production function defined by his ability (presumed exogenous), pedagogy offered him by the course, defined by teacher quality and the instructional method, e.g., self-paced instruction. Thus we can write:

$$(2)\quad \mathit{Achievement}^i = a + b_1^i\,\mathit{effort} + b_2^i\,\mathit{teacher\ quality} + b_3^i\,\mathit{SPI} + b_4^i\,\mathit{general\ ability} + b_5^i\,\mathit{analytic\ ability}$$

In retrospect, enjoyment of a course depends on his "profit," i.e., the difference between the grade from the course and the costs of the course $E^i$ weighted by psychic or financial cost of that effort to him. What he received will depend on his grades and his tastes. Thus our third equation of our model of the learning process became:

TABLE 1—ACHIEVEMENT

| | 1973–74 (Spring) | 1974–75 (Fall) | 1974–75 (Spring) |
|---|---|---|---|
| *General ability* | | | |
| *SAT* verbal | .0305 | .0248 | .0240 |
| (200–800) | (9.83) | (9.67) | (7.71) |
| *SAT* math | .0248 | .0278 | .0187 |
| (200–800) | (7.15) | (9.25) | (5.17) |
| Class rank | .0095 | .0087 | .0100 |
| (200–800) | (3.24) | (3.63) | (3.44) |
| Harvard scholar | 1.6397 | .6690 | 1.1969 |
| (1=yes, 0=no) | (2.24) | (1.06) | (1.57) |
| *Analytic ability* | | | |
| Uncomfortable with math | .0814 | .0454 | −1.4717 |
| (1=yes, 0=no) | (.007) | (.060) | (−1.48) |
| Sex (1=female) | −1.8644 | −1.027 | −1.9254 |
| | (−3.42) | (−2.48) | (−3.55) |
| Race (1=minority | .7068 | −1.1433 | −.9017 |
| | (.88) | (−1.77) | (−1.16) |
| Intended concentration (1=humanities) | −.3643 | −.8841 | −3.3623 |
| | (−.41) | (−1.15) | (−3.54) |
| Intrinsic interest | −1.0676 | −.2244 | −.4845 |
| (1=much, 7=none) | (−6.79) | (−1.77) | (−2.95) |
| *Pedagogy* | | | |
| Student's ratings of teaching fellows | .8057 | .0377 | −.0994 |
| (1=high, 7=low) | (4.64) | (.27) | (0.60) |
| Preparation hours (per week) | .1472 | NA | NA |
| | (4.48) | | |
| Teacher's grades in graduate school | .1525 | .4638 | .9594 |
| | (.61) | (2.67) | (4.72) |
| *SPI* | 3.8928 | 2.5161 | 3.0546 |
| | 4.45) | (2.83) | (2.54) |
| $R^2$ | .467 | .423 | .442 |

*Note:* The values in parentheses below the coefficients are $t$-ratios.

$$(3)\quad Enjoyment^i = a + b_1^i\, achievement + b_2^i\, effort + b_3^i\, tastes + b_4^i\, general\ happiness$$

Finally, believing that a continuing intellectual involvement with the discipline, as manifested by desire to take further economics courses is an important output of an introductory course for the student, we added a concentration equation:

$$(4)\quad Intended\ courses^i = a + b_1^i\, intended\ concentration + b_2^i\, intended\ career + b_4^i(Grade_{Ec\ 10} - grades\ in\ other\ courses) + b_5^i\, enjoyment + b_6^i\, SPI.$$

Thus, we had a model, admittedly crude and turning on heroic assumptions, which provided us with a framework and set of questions about *SPI* and the wherewithal to answer them; namely:

1. Does *SPI* have any net effect on learning, enjoyment or concentration? More precisely put, is $b^iSPI$ significant if all of the endogenous variables are suppressed?
2. How do these effects, to the extent that they exist, vary among students, i.e., does $b^iSPI$ differ systematically among definable classes of students?
3. If *SPI* has any effect on achievement, enjoyment or concentrations, what is the mechanism through which it works? Translated into the terms of our model the question becomes: Is $b^iSPI$ significant in the effort equation, but not in the achievement equation? Or vice-versa? Or in both?

## III. The Results

### A. The SP Effect

The answer to the first question is provided in Tables 1, 2 and 3. Table 1 shows the result of estimating equation 2 without the endogenous effort variable. Thus it provides an upper bound on the pure self-paced effect; Tables 2 and 3 show results for enjoyment and concentration equations, respectively. Given the space constraint and the similarity of results, only one semester's data is presented for the enjoyment and concentration equations.

Both the enjoyment and the concentration equations were somewhat disappointing. The *direct* contribution of *SPI* to student happiness is insignificant and its contribution to concentration decisions is small. In its initial year, 1972–73, *SPI* had produced both remarkably happy students and a very high proportion of concentrators. However, both the 1973–74

TABLE 2—ENJOYMENT (FALL 1974)

| | |
|---|---|
| *Achievement:* | |
| 1st semester grade | −.0909 |
| | (−3.87) |
| *Effort:* | |
| Percent reading not done | .0656 |
| | (1.85) |
| Attendance | .0519 |
| (1 = always, 7 = never) | (0.79) |
| *Taste variables:* | |
| Uncomfortable with math | .071 |
| | (.53) |
| Sex (1 = female | .6377 |
| | (3.80) |
| Athletic interests | .2029 |
| | (2.34) |
| Personalistic concerns | −.3062 |
| | (−1.99) |
| Religious interests | −.0959 |
| | (−2.66) |
| Entrepreneurial interests | .3268 |
| | (3.82) |
| Intellectual interests | .1557 |
| | (1.19) |
| *Pedagogy:* | |
| Student's ratings of teaching fellows | .3686 |
| (1 = high, 7 = low) | (7.45) |
| Teacher's grades in graduate school | .0681 |
| | (1.09) |
| *SPI* | −.1446 |
| | (−.70) |
| $R^2$ = .291 | |

*Note:* The values in parentheses below the coefficients are *t*-ratios.

TABLE 3—MORE COURSES IN ECONOMICS

| | |
|---|---|
| Intended to concentrate in economics | −1.0119 |
| (1=yes, 0=no) | (−3.00) |
| Intend law or business career | .4490 |
| (1=yes, 0=no) | (1.42) |
| *Taste variables:* | |
| Race | 1.0945 |
| (1=minority) | (2.12) |
| Sex (1=female) | −.3295 |
| | (.95) |
| Political interest | .7356 |
| | (4.01) |
| Personalistic interests | −.8525 |
| | (−2.81) |
| Athletic interests | .2682 |
| | (1.49) |
| Religious interests | .0970 |
| | (.84) |
| Entrepreneurial interests | .2440 |
| | (1.41) |
| Intellectual interests | .7159 |
| | (2.98) |
| Grade point average other courses | −.0039 |
| | (−.07) |
| Enjoyment | .7172 |
| (1=very much, 7=not at all) | (7.94) |
| Student's ratings of teaching fellows | −.1730 |
| | (−1.72) |
| *SPI* | .1029 |
| | (2.22) |
| $R^2=.422$ | |

*Note:* The values in parentheses below the coefficients are the *t*-ratios.

and 1974–75 results suggest very strongly that the 1972 experience was largely attributable to a Hawthorne effect of the very heavy faculty involvement in the grading process during the first year.

The achievement equation shows rather more encouraging results. On a multiple choice final made up of 10 *TUCE* questions, 20 *GRE* questions and 10 local questions with a 40 point maximum, and a mean score of 23, *SPI* does appear to improve performance by about 15 percent. Put differently, it substitutes for about 150 *SAT* points or 10 hours of weekly preparation by an instructor. Moreover, results available in the larger study mentioned earlier suggest that the differential increases after two years out of the course.

### *B. The Distribution of Benefits*

Having established that there were some identifiable education benefits associated with an *SPI* system it was appropriate to move to the second question: Who reaped those benefits? In particular, was the nonthreatening, repeat-play environment most helpful to the slower student and/or the bewildered humanities majors who did not yet understand how to study analytic material? Or was it an enrichment program which brought students into regular contact with very bright and stimulating upperclassmen and graduate students? To examine this question we stratified first by *SAT* scores (above 700/below 700), and then by year (freshman/nonfreshman).

Unfortunately the results were not as cleancut as we would have liked. (These regressions are available from the author on request.) In general, lower *SAT* groups and freshmen seem to be slightly more advantaged by *SPI* than higher *SAT* groups and upperclassmen. The differences between freshmen and upperclassmen are small and in any case not consistent over semesters. Thus it is clear we need to investigate the distribution effect with other group criteria. As an aside, the most interesting result of these regressions concerns the other instructional inputs: the marginal product of the well-prepared teacher is highest with the less brilliant student.

### *C. The SPI Technology*

Having established that a broad group of students did measurably better in a self-paced format, our next task was to find out why. The first hypothesis was that *SPI* had simply reinvented "study hall"; that *SPI* was primarily a device for getting students to work harder, or at least more steadily. To test this hypothesis we reestimated our achievement equation,

this time including the effort variables; homework hours, attendance, etc. The results for one semester, spring 1974–75, are shown in Table 4. Although the self-paced effect is slightly reduced it is not eliminated. This test is not conclusive. It is possible, for example, that *SPI* makes the study of economics so rewarding that hours seem like half-hours, and thus the reported work hours of the self-paced students are biased downward. Nevertheless, it appears *SPI* is not simply a system for sweating the students.

Given this finding we made one further test in an attempt to discover whether the defining of objectives, the self-knowledge, or the active interaction was the critical feature of SPI. For students in self-paced sections, we have two measures of participation: one, the number of self-paced exams each took and two, the number passed. (The number of *SPE*s attempted is generally larger both because many students failed a unit test on the first try, but also because after passing, some students would take additional tests "for practice.") Arguably, the number of *SPE*s passed measures self-knowledge, while number of *SPEs* attempted is a measure of the amount of interaction between student and grader. To discover which contributed more, we reestimated the achievement equation for the self-paced group only, using both *SPE*s attempted and *SPE*s passed as dependent variables.

TABLE 4—ACHIEVEMENT: FALL 1974
(Self-paced group only)

| | |
|---|---|
| *SAT* verbal | .0370 |
| (200–800) | (3.15) |
| *SAT* math | .0074 |
| (200–800) | (.75) |
| Class rank | −.0017 |
| (200–800) | (−.16) |
| Uncomfortable with math | .0221 |
| (1=yes, 0=no) | (.21) |
| Intrinsic interest | −.6103 |
| (1=much, 7=none) | (−1.48) |
| Sex | −.3958 |
| (1=female) | (−.29) |
| Race | 1.1903 |
| (1=minority) | (1.51) |
| Attendance | .4437 |
| (1=always, 7=never) | (.51) |
| Percent reading not done | −.6496 |
| | (−1.20) |
| Responded to questionnaire | −.6134 |
| (1=no/0=yes) | (.52) |
| Teaching fellow rating | −.5339 |
| (1=high, 7=low) | (−1.02) |
| *SPE*s passed | .5401 |
| | (3.51) |
| *SPE*s attempted | .2332 |
| | (2.14) |
| $R^2=.723$ | |

The rather remarkable results are shown in Table 4. "Self-paced exams attempted" captures a considerable amount of the variances, and even more interestingly, sharply reduces the coefficients on the ability variables. This result is consistent with the hypothesis that the crucial feature of *SPI* is the interaction between students and grader; that it is not just glorified homework but a very directed form of one-to-one instruction.

## IV. Some Tentative Conclusions

This study of *SPI* has no claims to be exhaustive: there are many other questions to be asked, and some better answers to be provided. Nevertheless, at least within the Harvard context it appears that we can purchase a fairly substantial and lasting improvement in scores for roughly $35 per student per semester. We seem to have allowed our students to substitute efficient, interactive hours for passive or unfocused hours although apparently not to their great pleasure.

Moreover, it is clear that moving to a *SPI* system brought about enormous improvements in our introductory course quite apart from its contribution to the learning of particular students. It forced us to reexamine our course objectives, to

# The Teacher Training Program for New Ph.D.s

*By* DARRELL R. LEWIS AND WILLIAM E. BECKER, JR.*

In 1970 the University of Minnesota Center for Economic Education initiated a teacher training program for graduate students in the Department of Economics. Prior to 1970, as is the case of most university departments of economics, new graduate students who were assigned to either the micro or macro principles course as section teachers were given a text, a sample syllabus, a class list, and a classroom number.[1] After five years of development and assistance from the Joint Council on Economic Education, graduate student instructors at Minnesota are now required to participate in a teacher training program. This paper is intended to provide an overview of the Minnesota program along with some preliminary evaluation results.[2]

## I. The Teacher Training Program

In their first quarter of teaching, graduate student instructors (*GSIs*) at Minnesota are expected to participate in nine weekly seminars dealing with teaching objectives, methods, and skills. They are also expected to participate in a three-part teaching performance feedback process which involves classroom videotaping, classroom test analysis, and student evaluations. In fact, participation in both the seminars and feedback procedures are now given as conditions of employment for all new graduate student instructors.[3]

### *A. The Seminars*

The seminar sequence of the program starts one week prior to the beginning of the fall quarter. The seminars are designed to acquaint the *GSIs* with the eleven teaching abilities identified in Table 1.[4] A senior faculty member together with an experienced senior graduate student takes responsibility for the content and manner in which the seminar is conducted. Participating *GSIs* are expected to complete basic exercises related to each of the eleven teaching abilities emphasized.

Seminar activities entail writing goals for the principles course, writing instructional objectives, discussing and demonstrating lecture skills, demonstrating alternative teaching techniques (large and small group), discussing and demonstrating the use of audiovisual materials, writing and grading examinations, evaluating student feedback, and developing student-

* Professor and Assistant Professor of Economic Education, respectively, University of Minnesota. We wish to recognize the significant contributions of Charles Orvis and Mike Salemi in the development and implementation of the programs described in this paper.

[1] On the basis of this observation the American Economic Association's Committee on Economic Education, in conjunction with the Joint Council on Economic Education, formulated its proposal to improve the teaching of college economics, G. L. Bach (1973).

[2] For a complete operational description of the University of Minnesota teacher training program in economics see Becker, *et al.* (1975).

[3] Since the pecuniary benefits to the graduate students are not weighted in favor of improving their teaching output, beginning in 1974 graduate students were required to participate in the teacher training program as a condition of employment, Becker (1975).

[4] Many of the instructional activities and pedagogical skills identified in Table 1 are discussed and developed in a forthcoming book edited by the Joint Council on Economic Education and entitled *Teacher Training Programs for Ph.D.'s in Economics*. This excellent source material is currently available for limited distribution in mimeograph form.

TABLE 1—TEACHING ABILITIES AND FEEDBACK

| Teaching Abilities Emphasized | Method of Feedback |
|---|---|
| 1. Recognize student needs | Student evaluations |
| 2. Set goals and write objectives | Daily teaching checklist |
| 3. Develop content sequencing | Student evaluation, videotaping, and teaching checklist |
| 4. Use lecture skills effectively | Student evaluations and videotaping |
| 5. Implement discussion skills effectively | Student evaluations and videotaping |
| 6. Demonstrate interpersonal communication skills | Student evaluations and videotaping |
| 7. Implement probing or questioning skills effectively | Student evaluations and videotaping |
| 8. Utilize multimedia materials effectively | Student evaluations and videotaping |
| 9. Construct and use well made testing instruments | Student evaluations and test item response |
| 10. Articulate problems and strengths in one's own teaching style | Student evaluations and videotaping |
| 11. Establish grading guidelines | Student evaluations |

teacher interaction and discussion techniques. Select economic education and educational literature is assigned to seminar participants as outside reading. All of this is intended to provide the *GSIs* with a working knowledge of pedagogical skills and techniques which they will hopefully try in their undergraduate classrooms.

### *B. The Feedback System*

A series of feedback processes are designed to reinforce the teaching abilities and skills presented to the *GSIs* in the seminars. These processes also give the *GSIs* an opportunity an analyze systematically their own teaching performance from the students' point of view. Each *GSI* is encouraged to assess his or her teaching performance on the basis of data collected from three feedback sources—i.e., a videotape observation system, student evaluations of the instructor via the Purdue Rating Scale for College Instructors, and test analysis aimed at evaluating content and application coverage.

The *GSIs* receive primary feedback on their actual teaching abilities by way of prearranged classroom videotapings of their instructional performances. Each *GSI* is videotaped in his or her undergraduate classroom at least three times during the fall quarter. The classroom videotaping of the *GSI* is coordinated by another experienced senior graduate student. This graduate student is also responsible for critiquing each *GSI's* videotape and assisting the *GSI* in spotting weak teaching practices which show up on the videotaped classroom presentation.[5]

To assist in reviewing and critiquing, two procedures and instruments were developed. Prior to each class designated for videotaping, the *GSI* completes a questionnaire directed to the objectives, content and techniques expected to be covered during the class period. This information is subsequently reviewed and compared with the videotape during the critiquing session.

The second instrument was constructed so as to measure actual instructor performance from the videotape. Prior to the review session with each instructor, the graduate student critiquer previews and codes each tape at twenty-second intervals according to a specially adapted observation scheme which measures (a) the

[5] Michael K. Salemi and Becker (1974) provide a videotape demonstration of the critiquing process.

method employed—lecture, question/problems, discussion, other, (b) the learning objectives—knowledge of facts, theoretical concepts, exposition on theory, simple application, complex application, and (c) the verbal and nonverbal expressions—supportive, receptive, neutral, unreceptive, disapproving. The data are then summarized and presented to the *GSI* during the review session.

The data from this latter instrument have proven valuable in at least two regards. Instructors, like students, respond to those things which are being measured. Secondly, the instrumentation is able measurably to confirm or reject those things the instructor says (thinks) he is doing in his classroom. It also reinforces the reviewer's intuitive critique and comments.

Using the standardized procedures and instruments developed, approximately two hours are spent with the individual instructor reviewing and critiquing his or her tape. This method of feedback gives the *GSIs* immediate peer reaction to their actual teaching performance.

The *GSIs* receive additional feedback on their teaching performance from student evaluations both during and at the end of each quarter of instruction. The 28 question Purdue Rating Scale for College Instructors is administered by the *GSIs* and processed by the Center for Economic Education. The printout given back to the individual *GSI* contains (1) a numerical student rating for each question, (2) a numerical student rating for each of five instructor characteristics measured by the Purdue form—personal, objectivity, exposition skills, tests and grading, subject matter knowledge—and (3) a range of average characteristic scores for all instructors over the past four years. The questions are all posed in such a fashion as to give implicit description for appropriate corrective action. This method of feedback gives the *GSIs* a comparable measure of their performance as perceived by students.

A third method of feedback which the *GSI* is taught to use is test analysis. The *GSIs* are taught in the seminars not only to use classroom tests to assess the achievement of their students but also to assist their own teaching performance. Each *GSI* is encouraged to use a variant of discriminant analysis to judge his or her teaching performance from classroom tests. In addition, the *GSIs* are taught to evaluate teaching output in terms of their cumulative students' test achievements, their colleagues' students' achievements and nationally normed student test results. This method of feedback gives the *GSIs* information which is directly related to the cognitive domain effects of their teaching.

Today, after five years of development, the seminar and three part feedback process are complementary inputs to the total teacher training program in the Department of Economics. Graduate students who are currently participating in the program may receive credit for their efforts under a graduate economics course number. The program is able to accommodate from six to fifteen new graduate students per year.

## II. Evaluation

Although there is no recent hard data on the effectiveness of the total training program described above, there is evidence from an earlier prototype program which suggests its effectiveness. Lewis and Charles C. Orvis (1973) carried out a controlled experiment in the 1971–72 school year to evaluate the impact of the initial University of Minnesota teacher training program.

The experiment used a control group of 323 fall quarter principles students (in 14 sections) and an experimental group

consisting of 438 matched winter quarter principles students (in 14 sections) with seven randomly selected and common graduate student instructors. The experiment was designed in such a way that the *GSIs* were not provided with any assistance or training during the fall term. However, during the winter quarter these same *GSIs* were systematically exposed to the department's training program.

Using ordinary least squares regression analysis, Lewis and Orvis included the standard educational regressors (age, sex, American College Testing Service score, grade point average, prior knowledge in economics, and student evaluations) with the student performance data collected. They reported that as a result of the training system student performance, as measured by the Test of Understanding in College Economics (*TUCE*), and instructor performance, as measured by the Purdue Rating Scale for College Instructors, both increased significantly. It was also found that the instructor performance ratings, as measured by student evaluations on the rating scale, associated highly with student performance on the *TUCE*.

## III. Summary

Writing off development costs as sunk cost (borne in part by special grants), the maintenance cost of the Minnesota teacher training program is not high. Variable cost is mainly dependent on the number and type of faculty and *GSIs* assigned to the program. At its present operational level with fifteen participants, only five major variable inputs are needed for maintaining and monitoring the system: (1) the senior faculty member to meet with the seminar and graduate student coordinators for approximately fifteen hours each quarter, (2) an undergraduate student with a quarter-time appointment for the videotape recording of the class sessions, (3) an experienced graduate student instructor with a third-time appointment in economics for scheduling, previewing, and critiquing each class session videotape, (4) an experienced graduate student with a third-time appointment for handling the logistics and demonstrations in the seminars, and (5) an eighth-time secretary to handle form processing and clerical responsibilities. These variable costs amount to about $9,000 per year. The fixed cost of videotape equipment is extremely low when amortized over a six-year period (about $500 per year). On a per graduate student basis, these costs are in line with other graduate course costs at the University of Minnesota.

The Department of Economics at the University of Minnesota has found the benefits to undergraduate students as well as to graduate student instructors outweigh the costs involved in the teacher training program. It is not surprising, therefore, that the history and geography departments at the University of Minnesota have recently adopted this program. Similarly, the graduate departments of economics at Florida State, Harvard, Indiana, Nebraska, Purdue, Wisconsin, and Yale have recently adopted forms of the program described in this paper. They have all joined the University of Minnesota in the Teacher Training Program for Ph.D.s in Economics sponsored by the Joint Council on Economic Education in cooperation with the Committee for Economic Education of the American Economic Association. Surely, other departments of economics and graduate schools can justify instituting similar teacher training programs.

## REFERENCES

**G. L. Bach**, "An Agenda for Improving the Teaching of Economics," *Amer. Econ. Rev. Proc.*, May 1973, *63*, 303–08.

W. E. Becker, "The University Professor as a Utility Maximizer and Producer of Learning, Research and Income," *J. Hum. Resources*, Winter 1975, 107–15.

W. E. Becker, D. Lewis, C. Orvis, R. Reizman, and M. Salemi, *A Training System for Graduate Student Instructors of Economics*, Univ. of Minnesota, Center for Educational Development 1975.

D. R. Lewis and C. C. Orvis, "A Training System for Graduate Student Instructors of Introductory Economics at the University of Minnesota," *J. Econ. Ed.*, Fall 1973, 38–49.

M. K. Salemi and W. E. Becker, "A Training System for Graduate Student Instructors of Introductory Economics," Univ. of Minnesota, Media Resource Center 1974.

# On Teaching Teachers to Teach

*By* STEPHEN H. LONG*

The purpose of this paper is to evaluate one implementation of the American Economic Association-Joint Council on Economic Education (*AEA-JCEE*) teacher training program. I shall briefly describe the University of Wisconsin version of the training program for graduate students, assess its strengths and weaknesses, and conclude with some thoughts about how we might alter our approach to teaching economists to teach.

The "Seminar on the Teaching of College Economics" began in the 1973–74 academic year. I was among the fifteen participants who joined the seminar which met once each week for one semester. Our studies encompassed course planning (including instructional objectives and task analysis), course implementation (lecturing, discussion techniques, and audio-visual materials), and evaluation of students and teachers (construction of tests, philosophy of grading, student evaluations of faculty, and critiques of video-tapes of seminar participants). The premise upon which the program was based was "whether or not great teachers are born (as opposed to made), most teachers can be trained to become better teachers." I believe this is true and there are some systematic data to support this proposition, in addition to the subjective evidence reported here.[1]

Interestingly, one of the most important benefits of the seminar had nothing to do with the content that has been described. Out of our studies and conversations came a kind of self-consciousness about the process of teaching. We became acutely aware that there existed alternative technologies for producing "Economics XYZ" as described in the catalogue of our first employer, ranging far beyond the choice of the appropriate text. Not only were we thinking about teaching, but we were talking about teaching—which for many of us was an important departure from the preoccupation with research fostered by graduate school curricula. One participant described this outcome as "consciousness raising." Some participants of the seminar are now, two years later, continuing to correspond by exchanging syllabi, exams, problem sets, and ideas about teaching.

A second important outcome results from the study of instructional objectives. There appears to be considerable agreement among the Wisconsin participants that this was the most useful seminar topic. Planning by stating goals in terms of demonstrable student outcomes focuses attention not on the content of the course, which is the stuff of graduate students' training, but on the object of instruction, the student. Commenting on the instructional objective approach, a former Wisconsin participant recently wrote: "I find myself constantly returning to questions

* Assistant Professor of Economics, Franklin and Marshall College. I am grateful to W. Lee Hansen for numerous discussions on the Wisconsin "Seminar on the Teaching of College Economics," and for sharing his 1973–74 students' written evaluations of the seminar. Jon Christianson, Robert Hutchens, Jack Mutti, and Thomas Pender kindly shared their thoughts on the seminar's influence on their teaching, two years after completing the course. Their comments have influenced this evaluation.

[1] Darrell R. Lewis and Charles C. Orvis (1973) present data on the Minnesota teacher training program which was effective in increasing pre-post *TUCE* differences and subjective student ratings of instructors.

of [instructional] objectives in planning courses, selecting textbooks, preparing lectures, etc."

Another significant impact of the program was the creation of a research strategy toward teaching. That is, our teaching is something to be studied and evaluated—not in a casual way (e.g., "I tried a new programmed text this semester and I think my students' essays were better")—but in a formal way, using all the expertise we bring to bear on other research questions. The seminar introduced us to the developing body of research on economic education and acquainted us with some evaluation instruments, including the Test of Understanding in College Economics (*TUCE*) and a number of forms for student evaluation of instructors.

Having developed these important general attitudes—that is, a self-consciousness about the teaching process, an appreciation of the merits of well-articulated instructional objectives, and a desire to carefully evaluate outcomes—the core material of the teacher training program proved useful in our teaching. Helpful tips on leading discussions, writing exams, and performing other instructional procedures were available from the reading materials or presented in class. But rather than describing these details, let me turn to some problems of the seminar.

All teachers must suffer from the problem of heterogeneous student backgrounds, and teaching teachers was no exception. Of the first seminar group of nine, six members were not currently teaching. Furthermore, six had previously taught or were currently serving as teaching assistants for a total of one semester or less. A tension developed between those who had taught more or who were not currently teaching and those who were teaching for the first time. The former exhibited a greater tolerance for abstract issues (e.g., philosophy of grading) while the latter were anxious for information that might be immediately applied in their classrooms (e.g., how to get a discussion started). A number of participants felt that a group having had the same amount of teaching experience would lead to better discussions. For those who had never taught, the seminar's role was primarily one of providing a perspective on teaching and sources for future reference, since many of the teaching skills required active practice to be understood and remembered. Another potential problem of our enrollment stemmed from the fact that participation was voluntary. It is likely that the participants were those students who were most interested in teaching. While they may have been the most receptive to training, they may also have needed it least among all the graduate students since they were more likely to pick up the material independently. Should we require participation by all graduate students, by all those who teach while in graduate school, or at least create incentives for participation?

Another concern of the participants was the nearly exclusive focus on the principles course. Some felt that examples and exercises should have been drawn from courses at the intermediate and advanced levels as well. The question here is whether or not there are important but different teaching problems at different levels. The teacher training program has been designed to focus attention on the various approaches or technologies for getting a specified body of material across to students. It emphasizes the principles course since it is most common to all participants and, some have argued, is the hardest economics course to teach.

The final source of difficulty was that some topics were given too much time, while others were ignored. Task analysis, a procedure for identifying the sequenc-

ing of material for presenting a concept, was seen as useless by a majority of the participants, at least as an improvement over common sense. Among the topics not covered in our seminar were choosing textbooks, managing teaching assistants, particular techniques for classes of varying size, showmanship, and treatment of academic dishonesty. After I moved to my present position in a liberal arts institution, I became increasingly aware of other issues which had been neglected, including the design of an undergraduate economics curriculum, the benefits and costs of team teaching, and broader philosophical issues of content. What occurs to me, however, is not that a program to train graduate students to teach can encompass all these issues, but that the appropriate model for teacher training may have two parts: (1) the seminar for graduate students, to provide the techniques most needed in the short run to "survive" the first few semesters of teaching, and (2) a follow-up seminar for previous participants two to four years out of graduate school, to treat the broader questions which are likely to arise only with experience.

Finally, I must mention that my title "On Teaching Teachers To Teach" purposely does not exclude the possibility of making the *AEA-JCEE* materials available to more experienced economists as well, perhaps through intensive summer workshops. As the demographics of academe lead to lower mobility, institutions may be expected to increase investment in the general teacher training of their older personnel. As a veteran of an *AEA-JCEE* seminar and an institution-sponsored faculty development program for persons of all disciplines, I strongly recommend that the profession continue its present efforts and that it extend their scope to include more experienced economists.

## REFERENCE

**D. R. Lewis and C. C. Orvis,** "A Training System for Graduate Student Instructors of Introductory Economics at the University of Minnesota," *J. Econ. Ed.*, Fall 1973, *5*, 38–46.

# One Participant's View of the Teacher Training Program

*By* Richard B. Hansen*

During the spring semester of 1974 the Department of Economics at the University of Nebraska offered a teaching seminar for graduate students which was based on a pilot Teacher Training Program sponsored by the Joint Council on Economic Education. When I was first asked to present my views as a participant in this seminar, I had mixed emotions concerning the insights I would be able to offer. I felt very strongly that I had benefited from the seminar and yet, in retrospect, it almost seemed that what I had been taught about teaching was common sense. Then, as I reflected on eight years of being a college student, I decided that most of what I had learned was, to use an analogy from economics, no more obvious than the fact that public debt differs from private debt, i.e., what I had learned was obvious only after it had been pointed out.

While I cannot furnish hard evidence, I will attempt to explain why I feel that the seminar was beneficial by relating some of the ways in which it influenced my teaching. My examples will be set in the context of the introductory economics course which served as the focal point of the seminar.

Our first seminar meeting dealt with the importance of establishing course goals and communicating them to the students. My own experience as a student reinforced the importance of this since I took my first economics course thinking that I was going to learn how to make money. As the course progressed, I got the same feeling I had had when I surreptitiously read *Lady Chatterly's Lover*—I kept wondering when I was going to get to the good part! In order to avoid this kind of misunderstanding, I tell my students on the first day of class that my primary goal is to help them become better informed citizens and that in order to attain this I want them to learn to apply economic theories and facts. Identifying what I want to accomplish in the course not only tells the students what to expect, it also helps me in selecting the topics that I cover and in determining how I treat those topics, e.g., I spend at least a full class period trying to convey the basic idea of economic efficiency and then ask the students to apply the concept to the present debate over the efficiency of pollution control devices.

A closely related point that was made clear in the seminar is the idea of telling the students what they are expected to learn. Formally, this entails the use of instructional objectives which, simply put, are specific statements of what the teacher wants the students to learn and how they are expected to demonstrate this learning. While I do not presently have time to construct properly written objectives, I do attempt to use the rationale behind them in several ways: first, by distributing study questions prior to each hour exam; second, by giving weekly quizzes covering material which I feel is important; and third, by posing questions in class which, as I tell the students, are indicative of

* Ph.D. candidate, University of Nebraska-Lincoln.

the type of question which I want them to be able to answer.

Another useful topic treated in the seminar was learning theory. In a very lucid and nontechnical article on the subject, G. L. Bach points out that learning involves the acquisition or modification of behavior so that students should be able to "do" economics if they have learned it. Additionally, Bach discusses the importance of motivation, relevance, clarity, feedback and reinforcement to the learning process. One strategy that I have developed which incorporates most of these elements, and with which I have had some success, is a weekly quiz. One class period in advance, I pose a question that requires the application of previously presented material to a contemporary or historical situation. By letting the students use what they have learned, this procedure demonstrates the applicability of economics (relevance) and encourages the students to learn the material (motivation); additionally, it provides repetition (reinforcement) and gives me the opportunity to let the students know how well they are doing (feedback) so that they can correct any deficiencies (or, as happened once, so that I can correct a deficiency in my presentation).

In terms of the number of ways in which my teaching was influenced, an article by Phillip Saunders and the accompanying seminar session on lecturing was dominant. Aside from convincing me that lecturing is a skill which can be developed, rather than an innate talent, the article provided many ideas which I have incorporated in my lectures. Let me mention a few. For one thing, I begin each class period by putting an outline of the material I want to cover on the blackboard. Second, after I have finished my presentation of a topic I usually pose a question which is intended to verify the student's comprehension of the material; additionally, this procedure gives me the opportunity to appraise my presentation and correct any inadequacies. Third, I often make notations to myself after class concerning aspects of the class that were especially good or poor so that I can make any indicated adjustments the following semester. Fourth, I have lost my reluctance to inject humor occasionally into my lectures (though, of course, humor is no substitute for preparation); this helps to put me at ease and, judging from my evaluations, creates an open atmosphere for the students. Finally, I have borrowed the idea of asking the students to complete an evaluation form after four weeks of class so that I can make any needed adjustments early in the semester (students indicate that they appreciate this).

Test construction is another area in which I feel more confident as a result of the teaching seminar. Of course, having fairly well-defined objectives in mind makes test construction a much easier task. And now that I am aware of the strengths and weaknesses of essay and multiple choice questions, I can more easily select the kind of question better suited to the topic and the level of understanding which I am testing. With respect to multiple choice questions, I have made use of a procedure called item analysis which reveals how many students selected each possible response and also indicates how students who performed well on the whole test fared on each question relative to those students who did poorly on the test. I have rewritten questions after discovering that only one or two students opted for a particular answer or finding out the "poor" students did better on a question than the "good" students.

In addition to these relatively specific examples, there are some broader, more general ways in which the seminar influenced my teaching. For one thing, I approach the introductory course as

though it were the only economics course the student would ever take; for me this has meant sacrificing theoretical rigor for realistic application. Second, I limit the amount of information which I cover in an effort to ensure that the students have a firm understanding of basic economics. Finally, and perhaps most importantly, I have become more interested in teaching as a result of taking the seminar; indeed, probably the most important lesson I learned came from an article by Ben Lewis which taught me to be unafraid of failure and unsatisfied with success.

It would be an error to conclude from what I have said that the teaching seminar was perfect. Considering that it was an inaugural venture when I participated, it is perhaps not surprising that there were shortcomings. For example, each participant was videotaped once in his classroom; since this is an excellent means of providing feedback to the novice instructor in his teaching, it seems to me that three or four tapings would be preferable (Nebraska now does three tapings). Student-teacher relations is another area to which I feel more time should have been devoted. Also, I think that the two hour session we had on audio-visual technique was too long (though certainly some familiarity with this technology could be useful).

In conclusion, I think that the Teaching Seminar was a worthwhile experience because it increased my interest in economic education and exposed me to techniques which I feel have improved my teaching. Considering the amount of time that our profession devotes to teaching a very important subject and the paucity of systematic effort generally given to improving instruction,[1] I would suggest that every Ph.D. granting department consider instituting this kind of program.

[1] The efforts of the Joint Council on Economic Education and the Committee on Economic Education of the American Economic Association represent important exceptions to this generalization.

REFERENCES

G. L. Bach, "What Should a Principles Course in Economics Be?" (undated mimeo).

B. Lewis, "A Retrospective Look at Undergraduate Economics," *Amer. Econ. Rev. Proc.,* May 1970, *60,* 370–75.

P. Saunders, "Lectures as an Instructional Method," (undated mimeo).

# INFORMATION AND MARKET STRUCTURE

## Information and Monopolistic Competition

*By* STEVE SALOP*

Information gathering and transmission have been gaining a larger role in the economy over time. M. Porat (1975) has estimated that the production, processing and distribution of information goods and services account for over a quarter of *GNP*. George Stigler (1961), George Akerlof (1970), and the papers in Edmund Phelps's volume (1970) showed the presence of imperfect information gives firms market power at least in the short run and often in the long run as well. Peter Diamond (1971) showed that the existence of costly information gathering led to an equilibrium in which each small firm charged not the competitive price but rather the monopoly price. Later papers have expanded Diamond's work for markets with different production and information gathering and transmission technologies. These papers find that if information is costly, each small firm obtains market power, and the equilibrium (if one exists) is characterized by prices above competitive levels and sometimes price dispersion as well. *The relevant market structure with imperfect information is not perfect competition but rather monopolistic competition.*

### I. The Basic Competitive Framework

There are a potentially infinite number of identical firms which can produce a homogeneous commodity from a production technology with fixed cost $T$ and increasing marginal cost function $MC(q)$. Thus, the average cost curve is U-shaped, attaining its minimum at $(p^*, q^*)$.

There are a large number of consumers $L$, who wish to consume this commodity. We begin by assuming that each consumer has an identical perfectly inelastic demand for one unit of the commodity at all prices below $p^m$. (Thus, $p^m$ can be interpreted as the dollar valuation of the commodity's marginal utility. It is most convenient to think of this commodity as a durable which is bought only once by each consumer.)

A monopolistic producer surely charges a price equal to $p^m$. If there is free entry into this market, new firms drive down the price until it falls to the competitive price $p^*$. No price dispersion is possible since each perfectly informed consumer will approach the lowest priced firm.

### II. Imperfect Information

In a Lancaster framework, the utility services derived from any commodity is a complex function of the characteristics that specify the commodity. The net utility surplus equals the utility services derived less its price. Given a choice among competing commodities, the "best buys" are those brands with highest net surplus.

Initially, a consumer has imperfect expectations regarding the net surplus of each good. These imperfect priors derive from his general knowledge and past information gathering and experience. It is useful to think of this as a dynamic learning process. The prior can be thought of as a probability distribution of the net surplus of a commodity. The distribution may

* Federal Reserve Board. These are the views of the author and should not be interpreted as the views of the Board of Governors. I am grateful to my collaborator, Joseph Stiglitz, and L. J. Dillard, M. Porat, and the editorial assistance of G. Edelen.

be an unbiased estimate of the actual net surplus (rational expectations) or may be biased due to incorrect extrapolation of the past, unseen commodity specification changes, past sampling realizations, false claims, or general ignorance. The formation of the prior is made more difficult since the relevant questions may not even be known until after some information is gathered.

The cost of information gathering will depend on the technology of information production and diffusion, the type, complexity and number of commodities, and the consumer's priors and preferences. These will define the speed of the learning process. Information gathering improves the quality of the posterior distribution. The expected value of information can be viewed as the increase in expected net surplus from the information. The effective cost of information includes its resource and nonpecuniary costs and will depend on the consumer's cost of time, preferences, analytic ability, experience and understanding of the market. For example, bargain hunters and (presumably) economists have low costs.

Depending on its cost, complete or incomplete information may be gathered. By complete information, we mean information that enables the consumer to purchase the best buy available. By incomplete, we mean that his expected net surplus will lie somewhere between the average and the best buy. For incomplete information gathering, the consumer must choose the sample size, minimum acceptable net surplus level, or reservation price.

Nelson (1970) contrasts two modes of information gathering, "experience" and "search." By "experience," we mean the knowledge obtained from actually using the commodity. This information may be gathered from personal use or from the experience of acquaintances (shared information). It is more likely to be utilized for inexpensive nondurables with unknown quality that are consumed each period. Experience information is generally incomplete, since every commodity is not sampled, and, in addition, the actual performance of a brand may be a random variable. By "search," we mean sampling before purchase by studying consumer magazines, visiting or phoning stores, reading advertisements, etc. Price information may be generated by search. In principle, search may yield complete information. For example, one could envision a magazine like *Consumer Reports* that publishes *all* the relevant information; a truthful guarantee has much the same effect.

There are important interactions (externalities) among consumers in the process. For example, if a consumer (having had one economics course) thinks that search by other consumers keeps the market honest so that price reflects quality, he need gather no information himself but will purchase the commodity with the greatest market share (John Conlisk and Dennis Smallwood). Alternatively, if a consumer believes in the Lemon's Principle, he will drop out of the market altogether (Akerlof).

Before analyzing this market, let us ask the question of why low cost consumers do not arbitrage the information differences away. Stated another way, since the marginal cost of distributing the information is nearly zero, why won't competition among information producers drive down the price? First, consumers face the same information problem with respect to information sellers as they do with respect to the commodities themselves. Each magazine must be evaluated and its content processed and analyzed. Furthermore, since the dissemination of the information by each magazine exhibits decreasing average costs, the information industry cannot be competitive in long-run equilibrium.

### III. Equilibrium

We now analyze the behavior and equilibrium price structure of this commodity market. Initially, the dynamics of the learning process are ignored. An equilibrium in this market is $n$-prices $p=(p_1, p_2, \ldots, p_n)$ each charged by a proportion $\beta=(\beta_1, \beta_2, \ldots, \beta_n)$ of the $N$-firms in production such that the following price and entry competition conditions are satisfied:

(1) *Price Competition* (*Maximum Profits*). Each firm chooses the price (and quantity) which maximizes its profits given its knowledge regarding the behavior of consumers and other firms (i.e., its demand curve).

(2) *Entry Competition* (*Zero Profits*). No new firm wishes to enter the market.

The exact specification of a firm's demand curve depends on the informational and strategy assumptions we make about firms. We will assume that a firm chooses a price given (i) the prices *chosen* by the other ($N$-1) firms and (ii) the search *rules* of consumers. Thus, a firm is a "Nash" competitor vis-à-vis other firms and a "Stackleberg" competitor vis-à-vis consumers. Condition (2) is the free entry assumption. We assume that entry occurs if profits are positive; hence, it is the zero profit condition.

In order to demonstrate the existence of the equilibrium, we follow the method of proposing a *potential* equilibrium set of $n$-prices $\hat{p}=\{\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n\}$, proportions $\hat{\beta}=(\hat{\beta}_1, \ldots, \hat{\beta}_n)$, outputs $\hat{q}=(\hat{q}_1, \hat{q}_2, \ldots, \hat{q}_n)$, and a number of firms-$\hat{N}$, and then see if any of the existing $\hat{N}$-firms wishes to change its price or if any potential entrants wish to enter. A potential equilibrium is an actual equilibrium if both deviant behavior and entry are unprofitable. An equilibrium can be pictured as in Figure 1.

The minimum price ($p_1$) will never be below the competitive price, for that would yield negative profits. Nor will the highest price be above the monopoly price. The exact specification of the equilibrium will depend on the production technology, the information gathering technology, and the joint distribution of consumers, utility valuations and search costs.

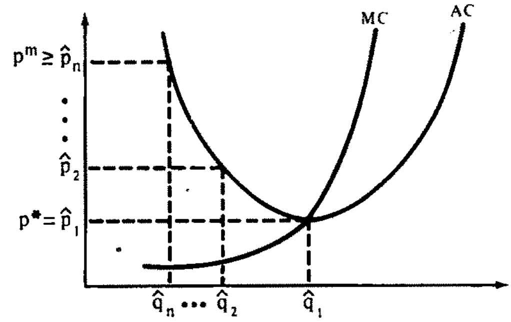



FIGURE 1

A simplified example can highlight the properties of the equilibrium. Assume that $L$-consumers all have identical valuations $p^m$, known with certainty. Prior price expectations are rational but limited. Consumers know the correct set of prices and proportions $\{p, \beta\}$, but a priori, they lack the knowledge of which firms charge which price. (This assumes away lemons and two-armed bandits.) For a cost $c_i$, consumer $i$ can purchase complete information from *Consumer Reports* in the form of the complete price-location correspondence. There are only two groups of consumers, a proportion $\alpha$ with cost $c_1 \geq 0$ and a proportion (1-$\alpha$) with higher cost $c_2 > c_1$. In spite of overwhelming empirical evidence to the contrary (A. Tversky), we assume consumers gather information in a rational, maximizing manner.

If both groups have perfect information ($c_1=c_2=0$), the competitive equilibrium ($p=p^*$) will obtain. If there are enough perfectly informed consumers ($c_1=0$, $\alpha$ high), the competitive equilibrium can still obtain in spite of $c_2>0$. This is an example of the externality informed consumers give to the uninformed. The weight of their search keeps the market competitive.

If both groups have positive cost, the competitive equilibrium cannot occur. A deviant firm could raise its price slightly above the competitive price and lose no customers. The only possible single-price equilibrium is the monopoly price, with entry competition driving profits to zero at this high price (Diamond). That equilibrium will obtain only if information costs are large for both groups. At this equilibrium, in order to induce consumers to pay the cost of discovering its location, a deviant firm would have to follow the unprofitable strategy of charging a price below the competitive price.

If lowering price to induce search is profitable, a two-price equilibrium (*TPE*) will obtain in which a proportion $\beta$ of the firms charge the competitive price ($p_l = p^*$) and the rest charge a higher price no greater than the monopoly price ($p^m \geq p_h > p^*$). Every type-1 consumer purchases complete information and pays the lower price. The higher price has the property that the type-2's do not find it worthwhile to purchase information. However, a proportion $\beta$ of the type-2's are lucky and randomly select a low-priced firm. Entry competition occurs until every firm makes zero profits.

When only complete information can be purchased, a *TPE* obtains even when there are many types of consumers (Salop and Stiglitz). In contrast, when incomplete information can be purchased, there will be as many prices in equilibrium as there are types of consumers (Gerald Butters, Peter von zur Muehlen).

If costs have little dispersion and are not too high, there is no *TPE* either. However, a single "limit price" $p^L$ exists with the following property. A deviant firm could increase its profits by raising its price above $p^L$. But if it does act in this short-run interest, it will destroy the equilibrium by inducing other firms to drastically cut price to induce search. Prices may begin to oscillate between the competitive price

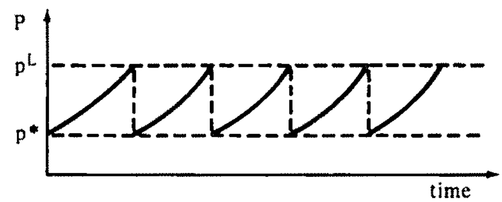



FIGURE 2

$p^*$ and the limit price. The exact dynamics will depend on the dynamic learning process by consumers and firms. In general, prices may creep up slowly to above the limit price which induces a price war down to $p^*$, only to again begin the upward creep. (See Figure 2.) The frequency and regularity of the cycle will depend on the time adjustment speeds of price changes, entry, and the learning by consumers.

An interesting example of this process is that of dynamically captive markets. Suppose consumers have better information regarding the brand they are now consuming than about other brands they consumed in the past or have never consumed. This asymmetry in information gives each firm market power over its current risk-averse customers since there are effectively "moving costs" of uncertainty from changing brands. If these moving costs are high enough, every firm can act like a complete monopolist over its segment of the market. Some deviant firm may find it profitable to subsidize these moving costs by charging a lower price. It gains customers who in turn become its captive as the information from their past experience decays. Hence, the deviant firm can creep its own price upwards towards the monopoly price.

Y. Shilony shows that a mixed strategy equilibrium in which each firm changes its price every period results from such a "captive market model," even when the commodities are nondurables bought every period. No firm's price ever settles down, so imperfections in information remain forever. This analysis also has the interesting feature that the firms could form a

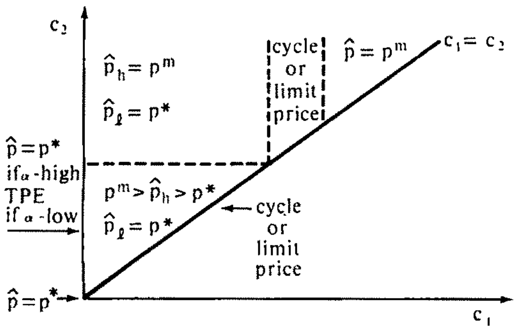



FIGURE 3

"noncooperative cartel" that needs to collude only once to set the random pricing rules. Once the cartel is begun, no single firm will ever find it profitable to cheat by deviating from the rule.

The various types of equilibria discussed are summarized in Figure 3.

### IV. Advertising

The firms themselves may engage in information transmission through truthful advertising and guarantees. This does not change the basic result. For example, suppose expensive large ads and inexpensive small ads are available, type-1 consumers are likely to notice both sizes, and type-2 consumers are likely to see only large ads. Then, firms that run large ads incur higher fixed advertising costs, charge a higher price and sell only to type-2's. On the other hand, suppose a firm sends ads out randomly, and each consumer chooses the firm with the cheapest advertised price from the ads received. In this model, price dispersion results even if consumers are identical (Butters). In both approaches, the highly advertised brands have higher costs and charge higher prices in equilibrium. Of course, this equilibrium result is different from the short run practice of a low priced deviant advertising its lower price. False advertising, bait-and-switch, dishonored warranties and other forms of fraud have also been assumed away in this analysis, though they are certainly profitable in the short run, and in the long run too if there are new suckers (ignorant consumers) born every minute.

### V. Price Discrimination

So far we have focused on price dispersion arising from search cost dispersion in which poorly informed consumers pay higher prices. When consumers differ in their valuations as well, monopolistic competition will generate price discrimination against consumers with both high costs and high valuations. Low valuation-high cost consumers will not purchase. A monopolist may also use dispersion in search costs to price discriminate against unidentifiable consumers with inelastic demand curves (high $p^m$'s here). If he permits price dispersion (through unadvertised specials or contrived heterogeneity), the high cost consumers will search less and hence pay higher effective prices on average. If these customers have more inelastic demand, then price dispersion acts as a price discriminating tie-in of search (a "bad") with the commodity he produces (Salop).

### VI. Competition vs. Welfare

The existence of many similar brands tends to increase effective search costs. Suppose a consumer must purchase a brand in order to gauge its utility (an "experience" good). Optimal search would consist of sampling brands *without replacement* until an acceptable one is found. In this case, entry competition has two countervailing effects. The usual entry effect increases choice and decreases prices. On the other hand, in sampling without replacement, small population sizes are preferred to large ones. For example, suppose there are 3 firms and a consumer evaluates the net surplus of these 3 brands as {5, 10, 15}. If he samples without replacement 3 times, he will certainly attain the maximum utility level of 15. If the number of firms doubles to 6 and the distribution of

utility levels remains identical (i.e., {5, 5, 10, 10, 15, 15}), then a sample of size 5 is required to obtain a level of 15 with certainty. A sample of size 3 will give expected maximum utility of only 13.1. Thus, the tradeoff between search intensity and utility level worsens, just as it would from an increase in effective search costs. In equilibrium, the increased cost tends to raise prices, *ceteris paribus*. The net effect of increased competition may be either to raise or lower prices. Thus, there may be a tradeoff between competition and static welfare.

## REFERENCES

**G. Akerlof**, "The Market for 'Lemons': Qualitative Uncertainty and the Market Mechanism," *Quart. J. Econ.*, 1970.

**G. Butters**, "Equilibrium Distributions of Sales and Advertising Prices," Princeton 1975, *Rev. Econ. Studies*, forthcoming.

**P. Diamond**, "A Model of Price Adjustment," *J. Econ. Theory*, 1971.

**P. Nelson**, "Information and Consumer Behavior," *J. Publ. Econ.*, 1970.

**E. Phelps**, et al., *Microeconomic Foundations of Inflation and Employment Theory*, New York 1970.

**M. Porat**, "The Information Economy," Inst. for Communication Research, Stanford 1975.

**M. Rothschild**, "Models of Market Organization with Imperfect Information," *J. Polit. Econ.*, 1973.

———, "A Two-Armed Bandit Theory of Market Pricing," *J. Econ. Theory*, 1974.

**S. Salop**, "The Noisy Monopolist: Imperfect Information, Price Dispersion and Price Discrimination," Federal Reserve Board 1973.

——— **and J. Stiglitz**, "Information, Price Dispersion and Monopolistic Competition," Federal Reserve Board 1975, *Rev. Econ. Studies*, forthcoming.

**Y. Shilony**, "Mixed Pricing in Locational Oligopoly," Berkeley 1975, *Rev. Econ. Studies*, forthcoming.

**D. Smallwood and J. Conlisk**, "Product Quality in Markets Where Consumers are Imperfectly Informed and Naive," Univ. of California, San Diego 1975.

**G. Stigler**, "The Economics of Information," *J. Polit. Econ.*, 1961.

**A. Tversky**, "Assessing Uncertainty," *J. Amer. Statist. Ass.*, 1974.

**P. von zur Muehlen**, "Sequential Search and Price Dispersion in Monopolistic Competition," Federal Reserve Board 1975, *Rev. Econ. Studies*, forthcoming.

# Information and Competitive Price Systems

By SANFORD J. GROSSMAN AND JOSEPH E. STIGLITZ*

Although the price system is conventionally praised as an efficient way of transmitting the information required to arrive at a Pareto optimal allocation of resources, the context in which the price system is usually discussed is not one in which the informational efficiency of the price system can be properly evaluated. Questions of how the price system leads the economy to respond to a new situation, how it conveys information from informed individuals to uninformed individuals, and how it aggregates the different information of different individuals, are never directly attacked.

In a series of papers (Grossman 1975a, 1975b, Grossman and Stiglitz 1975, and Stiglitz 1971, 1974), we have attempted to remedy this deficiency. It is the object of this paper to draw attention to some of the more fundamental implications of our approach and to use it to assess the meaning and validity of the efficient market hypothesis. Although our discussion will accordingly focus on the capital market, the kind of analysis developed here is applicable to any competitive market subject to random shocks.

## I. Prices and the Transfer of Information

The basic idea behind our analysis[1] may be illustrated by the following example:

Assume there are two assets, one safe and one risky, and that the return to the risky security $r$, depends on a random variable $\eta$, which can be observed at a cost, and another, unobservable random variable $\epsilon$:

$$r = \eta + \epsilon, \tag{1}$$

where $\eta$ and $\epsilon$ are independent, normally distributed random variables. Knowing $\eta$ reduces but does not eliminate the risk associated with the asset. The per capita demand, $X_I$, for the asset by those who are informed of $\eta$ will depend both on the price of the asset and the value of $\eta$.

$$X_I = X_I(p, \eta). \tag{2}$$

We assume that $\partial X_I/\partial\eta > 0$ and $\partial X_I/\partial p < 0$. Equilibrium each period requires that demand equal supply:

$$\lambda X_I(p, \eta) + (1 - \lambda)X_U(p) = X^s, \tag{3}$$

where $X_U$ is the per capita demand of the uninformed, $X^s$ is the per capita supply and $\lambda$ is the fraction of the individuals who are informed. Uninformed individuals observe only price, but from the price they may be able to infer $\eta$. For instance, if the stock of the resources were fixed, the uninformed individual can infer that a higher $p$ is associated with a higher $\eta$, since an increase in $\eta$ increases informed demand, and thus the price. Since there are no other stochastic elements in this model, there will be precisely one $\eta$ corresponding to any $p$. Hence, the conditional distribution of $r$ given $p$ is the same as the conditional distribution of $r$ given $\eta$. Thus, the price system conveys all the information from the informed individuals to the uninformed.

Now, let us introduce some further randomness; e.g., in the stock of the risky asset or in the demand functions of informed or uninformed individuals. Then the price

* This work was supported by National Science Foundation Grant SOC74-22182 at the Institute for Mathematical Studies in the Social Sciences (*IMSSS*), Stanford University. The authors are also indebted to the Dean Witter Foundation for financial support.

[1] See Grossman and Stiglitz (1975) for proofs and a detailed analysis of the model described by equations (1)-(4).

may be high because $\eta$ is high, but it may be high because the supply of the risky asset is low, or because informed individuals' demand functions have shifted upwards. Hence, corresponding to any $p$, there is a *distribution* of possible values of $\eta$. The price system conveys some information, but does not transmit all the information from the informed to the uninformed: on average, when the price is high, the return is high (i.e., $\eta$ and price are correlated) but the price is a *noisy signal;* that is $p$ and $\eta$ do not contain the same information about $r$.

Assume that the source of randomness is the supply of the risky asset. (We shall use this example through the rest of the paper.) Then, from (3), the equilibrium price will depend on $\eta$ and the stock of the risky asset, $X^s$; write $p=p(\eta, X^s)$. Solve for $\eta$ as a function of $(p, X^s)$ as, say, $\eta=\mathrm{t}(p, X^s)$. Using (1):

$$r = \mathrm{t}(p, X^s) + \epsilon. \tag{4}$$

The distribution of $(X^s, \epsilon)$ induces a distribution on $r$ for a given $p$. Since the uninformed observe $r$ and $p$, they come to learn the conditional distribution of $r$ given $p$. When they observe a $p$, they use this distribution to determine the expected utility from purchasing a given amount of the risky asset; $X_U$ is chosen to maximize expected utility. This is how the uninformed individual's demand function in (3) is derived. Finally for this to be an equilibrium, for all $\eta$ and $X^s$, $p=p(\eta, X^s)$ must be a solution to (3). Such an equilibrium entails rational, self-fulfilling expectations.

This is a reasonable condition for long-run equilibrium. If this condition is not satisfied (and the stochastic process describing the returns is stationary),[2] then an individual will eventually observe that the frequency distribution of returns, conditional on the observable variables, is different from the subjective distribution, and accordingly, ought to revise his expectations.

As there are costs of obtaining information, the marginal individual who chooses to become informed must be indifferent to being informed or uninformed, i.e., the increment in expected utility from becoming informed is exactly offset by the cost of the information. In making this calculation, individuals assume that a change in their information (and hence in their demands) would have no effect on prices. This is an adaptation to this context of the conventional Nash equilibrium hypothesis of competitive equilibrium theory.

Since when no one is informed, the price system conveys no information, the value of information about $\eta$ is likely to be high; when almost everyone is informed, the price system is very informative, so the value of knowing $\eta$ precisely is low. Thus, provided the costs of information are positive but not too high, equilibrium entails a fraction, $\lambda^*$, of the population being informed—that $\lambda$ which generates a price solution to (3) such that the marginal individual finds the expected utility to being informed equal to the expected utility of being uninformed.

Some striking features of the equilibrium which we have modeled should be noted. First, it provides a resolution of the following classical conundrum. If markets are perfectly arbitraged all the time, there are never any profits to be made from the activity of arbitrage. But then, how do arbitragers make money, particularly if there are costs associated with obtaining information about whether markets are

[2] One can argue that the limitation of our analysis to stationary stochastic processes is not a serious limitation; economic theory is concerned with identifying, describing, and explaining regularities in economic processes. Economic theory attempts to identify within a particular event those characteristics which it has in common with other events which have occurred. It is these regularities that are described by the stationary stochastic process.

already perfectly arbitraged? The conventional answer is that, when markets are not arbitraged, there are profits to be made, and so equilibrium *must* entail perfect arbitrage; the profits accrue in the process of responding to some unspecified disequilibrium. A particular example of this classical conundrum is presented by the efficient market hypothesis, which argues the prices on capital markets reflect all the relevant information instantaneously.

We resolve this paradox by arguing that there are constantly new shocks to the economy; although each of these shocks may have certain individual characteristics—the company president may be sick, a machine may break down—from the point of view of an analysis of market behavior, we are interested not in these individual characteristics, but in how these shocks affect market returns; and we postulate that we can describe the occurrence of these different shocks, in terms of their effects on returns, by a stationary stochastic process. The capital market must continually adjust to these shocks. We have formulated an equilibrium notion which explicitly takes account of the economy's response to these various shocks. Others have described this as a disequilibrium situation, but have been unable to say much about it.

In the structure we have developed, the market never fully adjusts. Prices never fully reflect all the information possessed by the informed individuals. Capital markets are not efficient, but the difference is just enough to provide the revenue required to compensate the informed for purchasing the information. The equilibrium fraction of informed traders $\lambda^*$ is determined jointly with the informativeness of the price system in such a way as to generate a competitive return to arbitrage.

Perfect arbitrage has one important implication—not all traders need to be informed. The informed traders make prices reflect true values, and the uninformed can simply take advantage of these services provided by the informed. In our analysis this is not true. Indeed, it is only because prices do not accurately represent the true worth of the securities (i.e., the information of the informed is not fully conveyed through the price system, to the uninformed) that the informed are able to earn a return to compensate them for the costs associated with the acquisition of the information.

Those empirical tests of the weak version "efficient market hypothesis" which show there are no gains to be made from looking at current prices and the past performance of the security provide support for our model, which assumes uninformed traders have rational expectations. But contrary to strong versions of the efficient market hypothesis, prices do not fully reflect all available information, in particular, that of the informed; the informed do a better job in allocating their portfolio than the uninformed. "Efficient markets" theorists state that costless information is a *sufficient* condition for prices to fully reflect all available information (Eugene Fama, p. 387). They are not aware that it is a necessary condition as well. But this is a *reductio ad absurdum*, since prices are important only when information is costly. (See Friedrich A. Hayek and Grossman 1975b.) Thus, an individual who throws darts at a dartboard to allocate his portfolio will not do as well as the informed individual;[3] what can be decided by a toss of the coin is not the allocation of the port-

[3] It is still true that if individuals were all identical and purchased the "market basket" of securities, the uninformed would do as well as the informed. Here we assume that the kind of information to make that feasible is not available. If individuals differ in their attitudes towards risk, or in their information structures, even when such a strategy is feasible, it may not be optimal.

folio but whether to be informed or uninformed.[4]

A second important characteristic of our analysis is that there is no proper separation between demand and supply. An increase in supply leads to a lowering of the price; since lower prices *on average* correspond to states in which returns are lower, the lowering of the price leads to a *lowering* of the evaluation of the risky security by the uninformed individuals, and hence of their demand. One cannot describe the equilibrium meaningfully in any period in terms of independently drawn demand and supply schedules, because the demand curves depend on the probability distribution of supply. This has the further consequence that an increase in price may actually increase demand; the presumption for a downward sloping demand curve is much weaker when individuals judge quality by price.

Still a third important and related observation is that prices, in our model, are serving two functions: not only are they being used to clear markets in the conventional way, but they convey information. In this sense, the models we have formulated are closely related to George Akerlof's lemons' model and to Akerlof (1973) and Stiglitz' (1975) analysis of labor markets.

The discussion so far has focused on the decision of whether to be informed or uninformed. There is an alternative way of looking at this question, which may shed some light on an old question discussed by John M. Keynes (p. 156). He suggested that the stock market might be viewed as a beauty contest, where the participants are not concerned with judging who is the most beautiful woman, but with judging who the other judges will believe is the most beautiful woman. Keynes made these remarks with more than a hint of disapproval; our analysis suggests that this may be unwarranted. It may be more efficient for some individuals to obtain information from others—through the price system or by other mechanisms—rather than obtain it directly.

## II. Prices as Aggregators

So far, we have discussed equilibrium in markets where prices convey information from the informed to the uninformed. In some market situations, different individuals have different information, and then the price system may serve to aggregate their information. That is, the demands for a risky security of an individual are affected by his information; total demand and accordingly equilibrium market prices thus depend on the information of all the individuals. In this sense the market price aggregates the various pieces of information.

A simple example may make this clearer. Assume there are a large number of isolated farmers. Each knows the size of his own crop, $y_i$. The size of the crop on any farm at any date is described by

$$y_i = \alpha + \epsilon_i \tag{5}$$

where $(\epsilon_i, \epsilon_j)$ are uncorrelated, $\alpha$ and $\epsilon_i$ are independent, normally distributed random variables with means $(\bar{\alpha}, 0)$ and variances $(\sigma_\alpha^2, \sigma_\epsilon^2)$, respectively. Thus, if $Y \equiv \sum_1^n y_j$, then $E(Y|y_i)$ is just a linear function of $y_i$, i.e., $E(Y|y_i) = h_1 + nh_2 y_i$.[5] Assume that there is a linear demand curve for the crop, so

$$Y = a - bP_s \tag{6}$$

where $P_s$ is the spot price next period. Then the subjective distribution of $P_s$ is normal, with mean $(a - E[Y|y_i])/b$ and a

[4] This is true only if no one has a comparative advantage in acquiring information.

[5] $E[Y|y_i] = n[\gamma\bar{\alpha} + (1-\gamma)y_i] \equiv h_1 + h_2 y_i$

where $\gamma \equiv \dfrac{\sigma_\epsilon^2(n-1)}{n(\sigma_\alpha^2 + \sigma_\epsilon^2)}$.

variance which is independent of $y_i$, $\sigma_p^2$. Since individuals differ in their expectations, there is an incentive to set up a futures market. Assume all individuals have constant absolute risk aversion, $k$. Then their demand for "futures" $Y_i^f$ is given by (where $P_f$ is the futures price):[6]

$$(7)\quad Y_i^f = \left[\frac{\dfrac{a - E(Y \mid y_i)}{b} - P_f}{k\sigma_p^2}\right] + y_i$$

and the market equilibrium requires

$$(8)\quad 0 = \sum Y_i^f = \frac{n}{k\sigma_p^2}\left\{\frac{a - h_1 - h_2 Y}{b} - P_f\right\} + Y.$$

Using (6), we obtain the result that the futures price is a linear function of the spot price:

$$P_f = h_3 + h_4 P_s.$$

It is a perfect aggregator of the information collected by the different individuals, i.e., by observing $P_f$, one can make a perfect prediction of the quantity available in the market and $P_s$.[7]

But there is a fundamental problem; if, as one would expect, individuals eventually come to realize that the futures price is a perfect predictor of the future spot price, then they will no longer base their demands on their own information, but rather base it solely on the market information. Since the futures price predicts the spot price perfectly (with zero variance) there is no need for hedging and there will be no trade. But without trade, there is no market; but without a market; their beliefs will differ. This paradox can be put another way. If the market aggregated their information perfectly, individuals' demands would not be based on their own information, but then, how would it be possible for markets to aggregate information perfectly?

So far, we have discussed some of the basic properties of our approach to equilibrium when information is costly. These models can also be used to address conventional questions related to existence, comparative statics, and welfare.

## III. Existence of Equilibrium Market Breakdown and Thinness

Both Akerlof (1970) and Grossman (1975a) argue that in markets where prices convey information between informed and uninformed traders, there is a possibility of market breakdown associated with a dwindling in the amount of trading. The example of the stock market presented above showed that this could indeed happen: if the price system were fully informative, there would be no differences in beliefs; and if there were no differences in beliefs, there would be no trade; but then it appears that it is prices in markets in which there are no trades which leads to uniformity of beliefs. Although this problem would be alleviated if prices did not perfectly convey information from the informed to the uninformed or if there were motivations for trade other than differences in information (e.g., differences in attitudes towards risk or in endowments), markets still might be thin, i.e., there would be a small volume of trade, and hence markets may be far from perfectly arbitraged.

Situations where markets might be thin or nonexistent need to be distinguished from those in which equilibrium does not exist. In the absence of noise, with costly information, an (Nash) equilibrium does

[6] Profits are $\pi = (P_f - P_s) + P_s y_i$. Then under normality and constant absolute risk aversion $k$, the individual maximizes

$$Y_i^f(P_f - E[p_s \mid y_i]) + E(P_s \mid y_i)y_i - \frac{k}{2}(y_i - Y_i^f)^2\sigma_p^2.$$

Solving for the optimal $Y_i^f$:

$$Y_i^f = y_i + \frac{E[P_s \mid y_i] - P_f}{k\sigma_p^2}.$$

[7] If $1 \neq nh_2/bk\sigma_p^2$, then $h_4 \neq 0$.

not exist,[8] since when one is informed, every individual believes he can become informed, increase his expected utility and not affect the market price. However, when a positive fraction of the population becomes informed the price system is fully informative, so it does not pay anyone to purchase the information.[9]

## IV. Welfare

The evaluation of the efficiency of the market in situations such as those analyzed in this paper is a subtle and difficult question. It is not obvious what the appropriate comparisons ought to be. Two alternative approaches might be delineated. In the *reformist* approach, we take as given the market structure, including the mechanisms for information transmittal. We ask simply, are there too many or too few informed individuals, or, is it desirable to have an information tax or subsidy? Although it is easy to show that the market solution is not, in general, efficient, it is difficult to ascertain whether there is too little or too much information acquisition. There are several effects, operating in different directions: some of the gains arising from *differential* information are private but not social returns, gains that some individuals make at the expense of others; on the other hand, since some information is conveyed by the price system, if that information is socially useful, those who purchase information generate a positive externality to those who do not. See Jerry R. Green (1973) and Stiglitz (1971). Even if there were no differential information, the price distribution does depend on the state of information. To return to our example of Section I, since when everyone is fully informed, price varies with $\eta$ and $X^s$, while when no one is informed, price varies only with $X^s$, it would not be surprising if information increases the variance of prices. Increased price variability is likely to lead to increased uncertainty about the value of one's endowments, and this is likely to lower expected utility. In one example we have analyzed in detail, where individuals have constant absolute risk aversion utility functions and randomly assigned endowments (all individuals having, however, the same endowment distribution), every one is better off if no one is informed than if all are informed.[10]

Finally, if the return to holding an asset for a period is the dividend plus the capital gain, the increased variability in price of the risky asset makes the risky asset riskier; thus, while in general, information reduces the riskiness of a risky asset, this is at least partially offset by this general equilibrium effect.

More fundamental questions are raised by the choice of alternative approaches to

[8] In the case where information is costless, an equilibrium exists; among the set of prices which might clear the market, there is a particular price function which clears it at zero trade and conveys all the relevant information, and this may be considered to be an equilibrium. There is no obvious mechanism for sustaining this particular set of prices, and this is a serious limitation.

[9] There could not exist an equilibrium in which trade occurred even if an individual had a monopoly power over information. For then the uninformed individuals would observe that they would do better not trading with the monopolist than trading with him, and the information-monopolist would simply determine equilibrium market prices. (See Stiglitz 1974.) Thus Jack Hirshleifer's classic analysis is not that of a competitive stock market with rational consumers. If his analysis refers to a market in which there is a monopolist in information, his results require irrationality on the part of other consumers in the market. If his analysis refers to a market in which the market for information-acquisition is competitive, then the results discussed in the text apply.

[10] This is a consequence of the unavailability of endowment insurance. This result has some important implications for a question which until now has not been satisfactorily resolved: Can there be destabilizing speculation? In this context, we interpret that to mean: Can the attempt to engage in intertemporal arbitrage lead to higher price variability which is associated with lower utility? The answer is yes, and indeed such attempts at intertemporal arbitrage can lower welfare. This occurs, in our constant absolute risk aversion example, because by the portfolio separation theorem, information has no allocative role.

information acquisition, e.g., a comparison between the decentralized process of the capital market and a centralized process. This, in some sense, was the central question of the Lange-Lerner-Taylor-Hayek debate.

Although this earlier debate was presumably about the informational efficiency of alternative organizational structures, models in which the systems had to adjust to new information were not formulated; rather it was argued that if the information were to be the same, the allocation would be the same, and thus, a comparison of alternative organizations came down to issues like a comparison of cost differentials arising from different patterns of information flows, or different speeds of convergence. Our analysis has suggested that a decentralized economy is likely to be characterized by individuals having *differential* information, that the separation in the earlier discussion of information and allocative questions is inappropriate, and that alternative informational structures will be characterized by different real allocations. In particular, Grossman (1975b) formalized Hayek's contention that prices are aggregators of information. There it was proved that if prices are sufficient statistics, the competitive economy where traders have diverse information generates allocations that cannot be improved upon by a central planner with all the information. However such markets do not provide incentives for information acquisition for the reasons given earlier. Thus only markets with noise will exist in equilibrium and these markets will not produce prices which are perfect aggregators. In this case a central planner with all the information can improve on the competitive equilibrium. Thus in our view the Lange-Lerner-Taylor-Hayek debate comes down to the fundamental distinction between economies where: (1) prices and hence allocations are the outcome of a competitive arbitrage process which will, of necessity, be imperfect because of the costs of arbitrage as discussed in this paper, and (2) economies where prices and hence allocations are the outcome of a centralized allocative mechanism which will, of necessity, be imperfect because of the costs of monitoring bureaucrats.

Thus, although we cannot provide an answer to whether a centralized or decentralized organization is more efficient, without more knowledge of the costs of operating a centralized informational mechanism, what we have established is that the conventional formulations of this question are misleading if not incorrect.

## REFERENCES

G. **Akerlof,** "The Market for 'Lemons': Qualitative Uncertainty and the Market Mechanism," *Quart. J. Econ.,* Aug. 1970, *89,* 488–500.

———, "A Theory of Information and Labor Markets," mimeo, Univ. of California, Berkeley, presented at the National Science Foundation-National Bureau of Economic Research Conference on the Economics of Information, Princeton 1973.

E. **Fama,** "Efficient Capital Markets: A Review of Theory and Empirical Work," *J. Fin., 25,* 383–417.

J. R. **Green,** "Information, Efficiency and Equilibrium," disc. pap. no. 284, Harvard Institute of Economic Research, Mar. 1973.

S. **Grossman,** "The Existence of Futures Markets, Noisy Rational Expectations, and Informational Externalities," technical report no. 182, *IMSSS,* Stanford Univ., Sept. 1975a.

———, "On the Efficiency of Competitive Stock Markets where Traders have Diverse Information," technical report no. 183, IMSSS, Stanford Univ., Sept. 1975b, *J. Fin.,* forthcoming, May 1976.

——— and **J. E. Stiglitz,** "On the Impossibility of Informationally Perfect Markets," paper presented to Dallas meetings of the Econometric Society, Dec. 1975.

F. A. **Hayek,** "The Use of Knowledge in Society," *Amer. Econ. Rev.*, Sept. 1945, *35*, 519–30.

J. **Hirshleifer,** "The Private and Social Value of Information and the Reward to Incentive Activity," *Amer. Econ. Rev.*, Sept. 1975, *61*, 562–74.

J. **Keynes,** *The General Theory of Employment, Interest and Money,* New York 1964.

J. **Stiglitz,** "Perfect and Imperfect Capital Markets," paper presented to New Orleans meeting of the Econometric Society, Dec. 1971.

———, "Information and Capital Markets," mimeo, Oxford Univ. 1974.

———, "Markets for Heterogeneous Labor with Imperfect Information," mimeo, Stanford Univ. 1975.

# Information, Screening and Human Capital

By John G. Riley*

Given the large and growing body of research into the nature and extent of human investment decisions, it is somewhat surprising that until the recent work of Michael Spence and Joseph Stiglitz there has been little discussion of the information transmission process. Certainly in all the theoretical modeling of human capital accumulation it has been implicitly assumed that throughout the life cycle employers are aware of each individual's marginal value product.[1] That is, "traditional" human capital theory has included the assumption that information costs are negligible.

However, it is by no means clear that a firm can evaluate cheaply the productivity of an individual worker, especially when the nature of the job is nonspecific (e.g., the management trainee). Plausibly information about an individual's value often unfolds only slowly with time on the job. Plausibly also, the costs associated with placing an individual in a job for which he is ill-suited are far from negligible. If so, firms have a strong incentive to offer salary contracts in which earnings are contingent upon long-run performance.

But such offers will only be completely successful in screening out the less productive if job seekers are either risk neutral, or have very tight prior probabilistic beliefs about their own lifetime productivity levels. Since it seems reasonable to reject both assumptions, contingent contracting of this type seems likely to be severely restricted.

An extension of this argument, emphasized by Stiglitz, suggests that firms are unlikely to incur large expenditures for on-the-job evaluations of their employees' potential. First of all, risk aversion makes job seekers unwilling to bear these costs in the form of considerably lower initial salaries. Secondly, the possibility of raids by other firms makes each firm unwilling to itself bear the costs of identifying top talent.

The question then arises as to whether firms might exploit information about an individual's general educational achievements at school and college in attempting to predict productivity on the job. Taking this to the extreme, might firms make initial job offers based entirely upon educational credentials which on average succeed in attracting workers of the desired productivity level?

The key element in answering this question is the manner in which marginal costs of education vary across individuals. With everyone staying in school until the marginal increase in earnings resulting from additional education is just offset by marginal costs, those whose costs are lower will plan to accumulate higher credentials. Then if the marginal cost of an additional unit of education is highly negatively correlated with productivity on the job, it will be the more skilled individuals who accumulate more education.

But it is surely reasonable to argue that the more productive workers are also, on average, the faster learners hence those with lower opportunity costs. Intuitively then, it seems quite possible that educational screening by firms might result in

* University of California, Los Angeles. The helpful comments of Sherwin Rosen, Michael Darby, Dennis de Tray, Jack Marshall and Finis Welch are gratefully acknowledged.

[1] For a brief summary of recent theoretical and empirical advances, see F. Welch.

a self-selection process through which individuals of differing talents are effectively separated. That is, although productivity is not directly observed until later in the career, educational achievement might "signal" information about this productivity to potential employees.

In the following section a model of human capital accumulation is developed which attempts to capture the essence of the above discussion as simply as possible, while retaining the basic structure of traditional human capital models. It is argued that despite the nonexistence of a Nash equilibrium there exists a unique set of wage offers which satisfies an appropriately weaker equilibrium concept. The section concludes with a very specific formulation of the model which highlights the problems of testing the educational screening hypothesis using the traditional earnings function approach. For this special model, a measure of the costs associated with screening is also derived.

## I. A Human Capital Model with Educational Screening

Consider a population of individuals differing in an innate unobservable characteristic $n$ which is distributed continuously over $[n_l, n_h]$. An individual of type $n$ can attain a level of education $z$ by staying in school for the first t years of his life.

$$\text{(1)} \qquad \mathrm{t} = \mathrm{t}(n, z) \qquad \mathrm{t}_n < 0, \ \mathrm{t}_z > 0$$

The variable $z$, observed by potential employers, is assumed to summarize an individual's educational achievements; classes taken, grade average, quality of schools attended and so on. The signs of the partial derivatives indicate that a higher $z$ represents greater educational achievement, and that an individual with higher $n$ is a faster learner. An additional restriction that we shall impose is $\mathrm{t}_{nz}<0$, that is, the time needed to achieve a marginal increase in education is everywhere lower for those with higher values of $n$.

While we shall treat $z$ as a continuous variable it is assumed that there are discontinuities in the underlying production of education. For instance, it seems reasonable to argue that as a first approximation an individual has a choice of colleges which are distributed continuously. However, having made a choice it then takes a discrete interval of time for the individual to obtain the desired credential. It is therefore difficult for a firm to obtain enough observations of t and $z$ to be able to infer an individual's value of $n$ prior to school completion.

Suppose further that the market value $M$, of individual productivity discounted to the time of entry into the work force t, depends both on the unobservable characteristic $n$ and upon education achieved prior to entering the labor market.[2]

$$\text{(2)} \qquad M = M(n, z) \qquad M_n, M_z \geq 0$$

We begin by considering a simplified "traditional" human capital model in which marginal productivity is costlessly observed. Assuming that the nontime costs of education are $c'$ per year, an individual staying in school for t years has a present value of lifetime income

$$P(n, z) = e^{-r\mathrm{t}}M(n, z) - \int_0^{\mathrm{t}} e^{-r\tau}c'\,d\tau$$
$$= e^{-r\mathrm{t}}\{M(n, z) + c\} - c$$
$$\text{where } c = \frac{c'}{r}$$

Since $P(n, z)$ is maximized if and only if the logarithm of $P(n, z)+c$ is maximized, the optimal education for type $n$ is the solution of

[2] It is simplest to think of $M$ as the discounted value of an individual's marginal productivity over all points in the life cycle, where the profile of the latter, for any given levels of $n$ and $z$ is determined exogenously. However, $M$ may also be interpreted as the solution of an on-the-job human capital optimization problem.

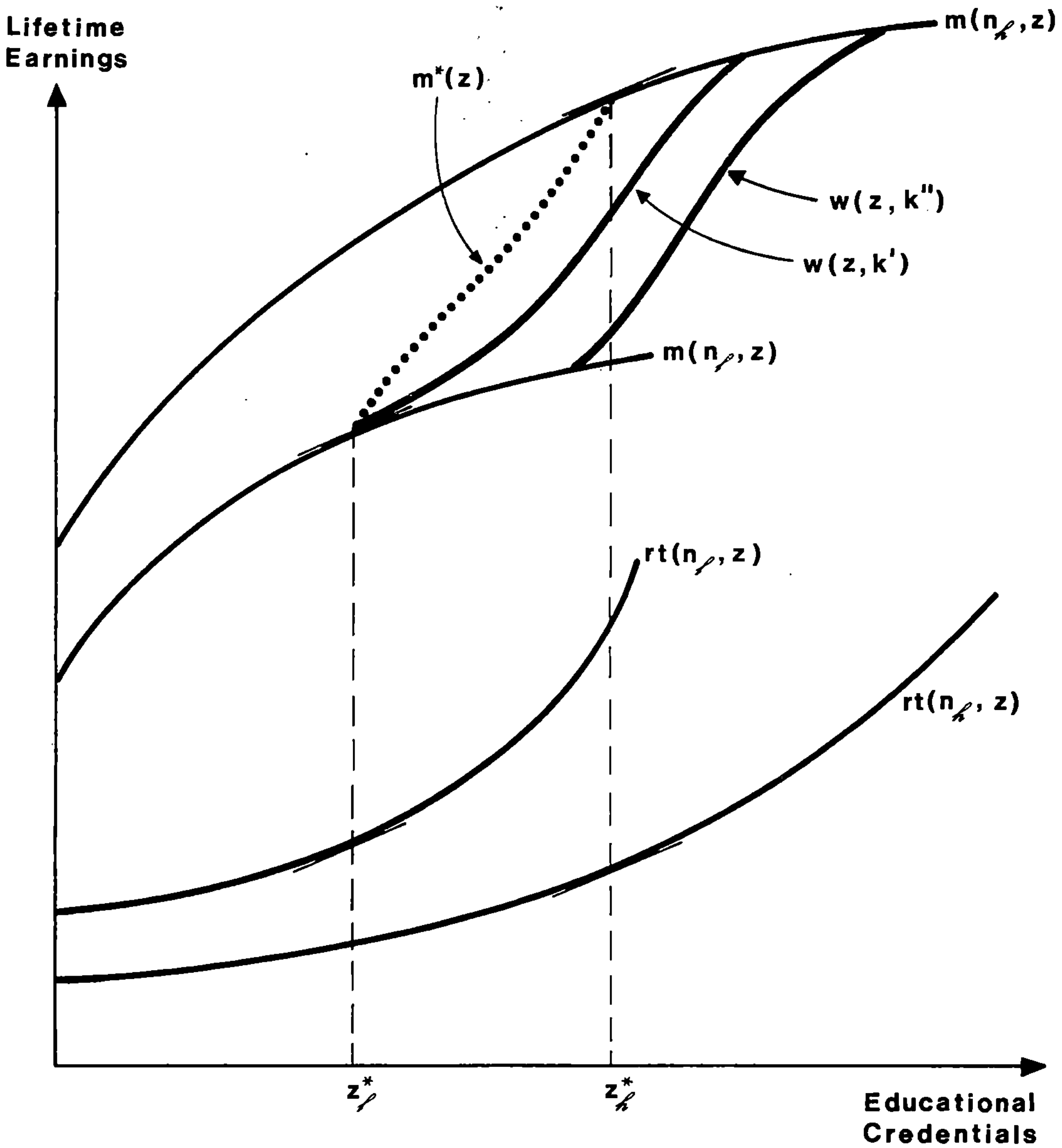



FIGURE 1: EARNINGS FUNCTIONS

$$\max_{z} \{m(n, z) - rt(n, z)\} \tag{3}$$

$$\text{where } m(n, z) = \log_e (M(n, z) + c)$$

The solution is depicted in Figure 1 for individuals at the extremes of the distribution of $n$. The curve $m^*(z)$ is the market equilibrium relationship between earnings and education. Making the reasonable assumption that the percentage increase in productivity associated with an additional unit of education is greater for those with higher $n$ (implying $m_{nz}>0$), it can be readily shown that optimal education $z^*(n)$ is strictly increasing in $n$.

We now consider the screenist view that productivity on the job is very costly to measure, at least in the early years. Sup-

pose each individual were to be offered discounted earnings $W(z)$ based on his educational achievements summarized here by $z$. Writing $w(z)=log_e(W(z)+c)$ and arguing exactly as above, the optimal choice of type $n$ is the solution of

$$\text{(3')} \qquad \operatorname*{Max}_z \{w(z) - rt(n, z)\}$$

Differentiating (3)′ twice yields the necessary conditions

$$\text{(4)} \qquad w'(z) - rt_z = 0 \qquad z > 0$$

$$\text{(5)} \qquad w''(z) - rt_{zz} \leq 0$$

Furthermore, differentiating (4) with respect to $n$ yields

$$(w''(z) - rt_{zz})\frac{dz^s}{dn} - rt_{nz} = 0$$

By assumption $t_{nz}$ is negative hence

$$\frac{dz^s}{dn} > 0,$$

that is, *those more favorably endowed with the unobservable $n$ attain higher educational credentials.*

Of course (4) and (5) are not sufficient for an equilibrium. What is required in addition is that employers' predictions about productivity levels should eventually be confirmed. Then given competitive behavior by firms, a further condition for equilibrium is that the gross earnings of those choosing an educational level $\hat{z}$ must equal their marginal value product.

$$\text{(6)} \qquad W(z) = M(n, z)$$

or, equivalently

$$\text{(6')} \qquad w(z) = m(n, z).$$

Given our assumptions about the functional forms one can in principle eliminate $n$ from (4) and (6′) and thereby obtain a first order differential equation, $w'(z)=f(w, z)$. Since solving such an equation yields an undetermined constant of integration, there are a whole family of earnings profiles $w(z, k)$, satisfying the following Spencian equilibrium conditions.

N–1. Individuals choose that education which maximizes discounted earnings.

N–2. The earnings of individuals achieving a given level of education equals (the average of) their marginal value products.

Two such profiles are depicted in Figure 1. It is easy to check that individuals accumulate more education if costs of observation result in its use as a screen. Differentiating (6′) totally with respect to $n$ yields

$$[w'(z) - m_z(n, z)]\frac{dz^s}{dn} = m_n.$$

Since we have already shown that

$$\frac{dz^s}{dn} > 0,$$

we have

$$\begin{aligned}\text{(7)} \quad m_z(n, z^s(n)) - t_z(n, z^s(n)) &\\ \leq w'(z^s(n)) - rt(n, z^s(n)) &\\ = 0 \qquad \text{from (4).}&\end{aligned}$$

But if $m$ were costlessly observable individuals would choose $z$ in order to make the left-hand side of (7) equal to zero. Then $z^*(n) \leq z^s(n)$ as depicted.

Spence's initial interpretation of these results was that an economy might settle in any of the states characterized by the arbitrary constant of integration $k$. In his terminology the conditions N–1 and N–2 yielded a *family* of potential "signalling" or "informational" equilibria. However the options open to firms were never adequately explored. It has since been shown by the writer that with simple extrapolative experimentation, all but one of the equilibria begin to unravel. The one profile that survives such experimentation, $w(z, k')$, has the property that those most disadvantaged ($n=n_l$) receive exactly the same earnings as they would if the productivity could be costlessly observed.

There is an appealing intuitive argument as to why this should be the case. In contrast with those elsewhere in the distribution, those on the "bottom rung" have no one lower down the ladder from whom they must be distinguished in order to earn a higher income.

The question as to whether this remaining earnings profile is an equilibrium is however a delicate one. As Rothschild and Stiglitz have indicated, there are circumstances in which there exists *no* set of contracts [in our case an earnings profile $w(z)$] satisfying N–1, N–2 and the following requirement for a Nash equilibrium.

N–3. Every alternative contract (that is, an earnings offer based on educational credentials) that might be offered by a single firm is nonprofitable.

Since it can be shown that for the model derived above, $w(z, k')$ does not satisfy N–3, it is perhaps tempting to reject this screening model on theoretical grounds.

However, N–3 seems unnecessarily strong since any firm exploiting it must also fear the reaction of other firms. This idea is formalized in the following weaker "reactive equilibrium" condition.

N–3′. For every alternative contract which is profitable if offered by a single firm, there is a second contract which yields profits to a responding firm and losses to the first firm.

It should be noted that arbitrary threats are excluded by the requirement that the revised contracts of a second firm must yield profits to that firm. This seems an appropriate restriction in the context of large numbers of competing firms.

Given that in the skilled labor market individuals are very often implicitly signing long-term contracts, it is valuable for each firm to develop a reputation about its salary structure, the opportunities for on-the-job training, etc. It therefore seems unlikely that any one firm could make a "quick killing" by switching to an alternative offer that yielded only short-run gains. Then at least for our application of the screening model N–3′ seems sufficiently strong to imply stability. Since $w(z, k')$ can be shown to be the unique profile satisfying N–1, N–2, and N–3′, it seems reasonable to describe this as the competitive screening equilibrium.

At this point it should be noted that considerable weight has been placed on the idea that productivity is a function of educational achievement, $z$, and that employers observe $z$. However, because of data limitations, most empirical studies have focused upon the return to years of education t.

While for the competitive screening equilibrium it is not necessarily true that years of education are an increasing function of $n$, we shall see for a specific example that this is indeed the case. Suppose $\mathrm{t}(n, z)=n^{-\alpha}z^{\beta}$ and $M(n, z)=nz^{\gamma}-c$. Then combining the necessary conditions (3′) and (6′) it can be shown that the equilibrium earnings profile has the following form

$$(W+c)^{\alpha}=k'+\frac{r\beta\alpha}{\beta+\alpha\gamma}z^{\beta+\alpha\gamma}$$

where $k'$ is determined by the requirement that those at the bottom of the distribution are exactly as well off as if $m$ were freely observable.

Eliminating $z$ and taking logarithms then yields

$$\begin{aligned} w &= log_e (W+c) \\ &= \alpha^{-1} log_e k' - \alpha^{-1} log_e \left(1-\frac{r\beta\alpha}{\beta+\alpha\gamma}\mathrm{t}\right) \\ &\approx A + Br\mathrm{t}; \qquad B>1 \end{aligned}$$

Of course this is precisely the Mincerian "earnings function" of traditional human capital theory. Since $B$ exceeds unity for all values of the parameters, our screening model also predicts a "rate of return" to

years of schooling in excess of the rate of interest. Clearly then, estimation of a single log earnings function will not yield inferences about the importance of educational screening.

We conclude this section by asking how costly is the screening mechanism as a method of predicting productivity differences. All that can be said for the general case is that everyone except the most disadvantaged ($n=n_t$) spend more time in school than he or she would if productivity were costlessly observable. Using the specific functional forms introduced above we can go further and ask just how much worse off individuals are in the screening equilibrium.

For simplicity we consider only the case in which nontime costs of education are negligible ($c=0$). Then in a world in which productivity could be costlessly observed, each individual would choose an education level and number of years in school to maximize $e^{-rt}nz^{\gamma}$, where $t=n^{-\alpha}z^{\beta}$. It is easy to check that for this special case every individual would choose the same number of years of schooling $t^*=(\gamma/r\beta)$.

Then writing the discounted earnings in the free information world as $V^*$ and the discounted earnings in the screening world as $V^s$ we have:

$$V^*/V^s = e^{-rt^*}m(n, z^*)/e^{-rt^s}m(n, z^s).$$

Substituting for $z$ using the special functional forms yields:

$$V^*/V^s = e^{r(t^s-t^*)}(t^s/t^*)^{-rt^*}. \tag{8}$$

Since in the screening equilibrium those at the bottom of the distribution are no worse off, we have $t^s(n_t)=t^*$. The "cost" of educational screening can then be calculated under different assumptions about the lower bound of the screened distribution. Assuming that only those going to college are screened ($t^*=12$), the cost of screening is hardly insignificant. With an interest rate of 6 percent the ratio $V^*/V^s$ climbs to 1.03 for those completing four years of college. After six years the ratio rises to 1.07.

Of course any implications drawn from one simple model must remain highly speculative. However it seems fair to conclude that the screening process described above is a rather costly method of differentiating highly productive workers from the rest of the pack.

## II. Concluding Remarks

The primary goal of this paper has been to link up the recent theoretical modeling of information[6] transmission via self-selection, with the traditional human capital literature. In examining a simple model of human capital accumulation we have suggested that there exists an appropriate definition of a competitive equilibrium in which individuals are hired on the basis of their credentials and are paid wages equal to actual productivity.

While the focus has been on the educational development of individuals prior to their accepting employment, it should not be inferred that on-the-job training is to be regarded as secondary. The screenist position is that education both enhances individuals' productivity levels and provides a mechanism for sorting out differences in these levels. Then one of the crucial roles of educational prescreening is to allow employers to select the more talented for jobs which involve considerable on-the-job training. Unfortunately this has not been recognized in several recent empirical studies which attempt to shed light on the educational screening hypothesis. As a result their conclusions should be regarded with considerable caution.[3] Indeed given that the Mincerian semi-log earnings function is derivable

[3] A major problem with these papers is that the "tests" are in the main not derived from an explicit model of the screening process. Where such modelling is implicit, it invariably involves a highly simplistic interpretation of the screenist position.

from a screening model, it might be tempting to infer that the educational screening hypothesis is not testable.

However, to take such a position it must be argued that direct monitoring of job performance is very costly in all skilled occupations. Surely it is more plausible that screening, if important at all, will be much more so for some classes of jobs than for others. Elsewhere in Riley (1976), it is shown that theory does yield predictions about differences between earnings functions for a class of jobs in which screening is used, and for a class in which productivity is cheaply monitored. These predictions are shown to be consistent with an empirical investigation by Kenneth Wolpin. While the latter draws a negative conclusion, it is argued that his results are more appropriately reinterpreted as providing mild support for the educational screening hypothesis. However, stronger conclusions must await a more systematic examination of the evidence. Hopefully future empirical work will reflect a clearer understanding of the potential dual role of education as both a mechanism for transmitting information about job applicants and for increasing productivity.

REFERENCES

**J. Albrecht,** "The Use of Educational Information by Employers," paper delivered at the Winter 1974 meeting of the Econometric Society.

**K. J. Arrow,** "Higher Education as a Filter," *J. Publ. Econ.*, 1973, *2*.

**R. Layard and G. Psacharopoulos,** "The Screening Hypothesis and the Returns to Education," *J. Polit. Econ.*, Sept. 1974, *82*.

**J. Mincer,** *Schooling, Experience and Earnings*, National Bureau of Economic Research, New York 1974.

**J. G. Riley,** "Competitive Signalling," *J. Econ. Theory*, 1975, *10*, 175–86.

———, "Towards a Test of the Educational Hypothesis," Univ. of California, Los Angeles, disc. pap., Jan. 1976.

**M. Rothschild and J. E. Stiglitz,** "Equilibrium in Competitive Insurance Markets: The Economics of Imperfect Information," *Quart. J. Econ.*, forthcoming.

**A. M. Spence,** *Market Signalling: Information Transfer in Hiring and Related Processes*, Cambridge, Mass. 1974.

**J. E. Stiglitz,** "The Theory of Screening, Education and the Distribution of Income," *Amer. Econ. Rev.*, June 1975, *65*, 283–300.

**P. Taubman and T. Wales,** "Higher Education: An Investment and a Screening Device," in *Education, Income and Human Behaviour*, New York 1973.

**F. Welch,** "Human Capital Theory: Education, Discrimination and Life Cycles," *Amer. Econ. Rev. Proc.*, May 1975, *65*, 63–73.

**K. Wolpin,** "Education and Screening," doctoral dissertation, City Univ. of New York 1975.

# Price Signaling in Experimental Oligopoly

*By* AUSTIN C. HOGGATT, JAMES W. FRIEDMAN AND SHLOMO GILL*

## I. Introduction

We report here a few results of oligopoly experiments which we carried out several years ago. These are laboratory experiments with human subjects along the lines of Lawrence Fouraker and Sidney Siegel (1963), Friedman (1967) and Hoggatt (1969). We sought to observe the predictive power of various theoretical equilibria. Then, to the extent that theoretical equilibria do not predict observed behavior, we seek to understand that behavior as best we can in the hope that the existing body of theoretical knowledge can be improved upon.

We used a simple textbook model of an oligopoly, then arranged for undergraduate student subjects to be the "firms" in the market. Our markets have 2, 3, 4 or 6 firms, each represented by one subject. One single "game" consists of a period of 1½ to 2 hours during which one group of subjects forms a market. For example, in a 3-firm market there would be 3 subjects each in a different room in the laboratory. Each would make the price and output decisions for his firm, and the pay to each for his participation in the game would consist only of the profits of his firm. At the beginning of the game, each would make the price and output decisions for his firm, and the pay to each for his participation in the game would consist only of the profits of his firm. At the beginning of the game, each subject would make his initial price and output decision. These initial decisions would remain in effect until changed by a subject; and he would be free to change either his price or output (or both) at any time. A particular game or market with its fixed set of firms (subjects) would remain in continuous operation for approximately an hour and a half.

The time periods are approximately 40 seconds each; and every fourth period, the subject receives a report on his sales, production, inventory, profit, etc. These time periods are sufficiently short that they approximate continuous time for the subject, especially given that there is not time in the whole game, except the beginning, when the subject *must* actively choose a price and an output level. Anytime afterward he can change his price and/or output level by entering new ones. They take effect at the start of the time period following.

Our experimental design was chosen with several objectives in mind: casual evidence, formal evidence (Friedman 1967), and intuition all suggest that sub-

* Professor of business administration, University of California, Berkeley, Professor of economics, University of Rochester, and Research Assistant, Center for Research in Management Science, University of California, Berkeley, respectively. We are grateful for research support to the National Science Foundation for GS-2423 and GS-2463, to the University of Rochester and to the University of California. The Center for Research in Management Science, Berkeley, where this research was carried out, operates a computer controlled laboratory for the study of human behavior. A manual describing the experimental control programs and the observations on which this paper is based is available on request.

jects with no prior experience in the experimental setting will behave differently than they will after getting experience. We opted for experienced subjects; hence, we provided the experience. Further, we wanted reliable subjects who would come to the laboratory when they were signed up and we wanted the experiment to be a relatively large part of the subjects' activities over the days it was in progress. Meeting these objectives required: (1) recruiting a very large number of subjects who played several (three) games each which were preliminary to the main experiments. These games allowed a chance for subjects to drop out before doing so caused us problems, and also gave the subjects the necessary prior experience; (2) choosing from the subjects who remained through the three preliminary games three groups of 9 each who would be in three separate experiments. Each group of 9 had to have large blocks of available time in common so that the 8 games of each experiment could take place within a few weeks. These 8 games consisted of 2, 3, 4 and 6 firm symmetric (i.e., same cost functions for each firm) and asymmetric games.

The three groups of 9 subjects played under differing information conditions. The first (A group) had "complete information" and "forecasting." By complete information we mean that each subject knew the demand and cost parameters for all firms in his market. Under "incomplete information" he would know only his own. By forecasting is meant a sequence of time periods before the game begins during which the subjects make choices, as usual, but the periods are for practice. They do not gain or lose money. This was done by Fouraker and Siegel (1963) in some of their experiments. The other two groups operated with incomplete information with forecasting (B group) and complete information without forecasting (C group).

A final design characteristic of great importance is that the games are noncooperative and the players were unable to communicate with one another by means of written messages, oral conversation, etc. By "noncooperative games" we mean that the subjects could not make enforceable agreements.

## II. A Summary of Some of the Results

Signals are special price choices which are chosen to substitute for verbal communication when the latter is impossible. All price choices could be defined to be communication, but it may not be fruitful to do so. Most price choices are likely to be routine and dictated by simple and straightforward (relatively short run) profit considerations. Furthermore, the routine nature of most price choices is likely to make them pass without much notice. A very unusual price choice, by contrast, is likely to stand out and to cause other players to wonder and to seek specific explanations. One explanation is that the chooser of the unusual price is trying to say something to the others through his action. Given the shortness of the time periods, it is possible to suffer a low rate of profit for a short time to deliver a message and yet not bear much cost.

We have tried to look at signals in the overall context of price policy. In this section, some of our results on price policy, including those related to signals, are sketched. Production policy is left out of account because it appears to have no significant influence on price policy.[1]

To get a complete picture of price behavior, several separate pieces of analysis must be put together. The first is whether to increase price, leave it unchanged or decrease it. We estimate probabilities for

[1] In our report, we go into production policy in detail.

each of these as a function of variables of interest. Second, if price is changed, is the change to be a signal? Finally, if a routine price change is to be made, what should be the magnitude? These topics are taken up below.

### A. The Definition and Identification of a Signal

We recognize two distinct types of "unusual" price choice in the data. The first is a price change of unusually large magnitude. We call these "steps," and they may be of either sign. Their identification is deferred for a moment. The second is a sequence of two or three successive price changes which sum to zero. The simplest example is an increase in price of $dp$, followed in the next period by a decrease of the same magnitude. We have named these "pulses." Pulses are designated with the sign of the first price movement, making the pulse in the example positive. The crucial element of the pulse is that within three periods, price has returned to where it began. Pulses may be unambiguously recognized from the definition given above; however, steps cannot because no criterion has been introduced to separate them from more ordinary price changes.

We obtain a sample of observations on price changes and other variables by first deleting all periods in which price change is zero. Second, we delete all periods which are the first of the game because these may be shown to come from a different population. Third, we delete all observations which are part of pulses. The remaining observations are divided into two subsamples, one for price increases, the other for decreases. For each of the 8 types of game, using ordinary least squares we regressed positive price changes on the following variables: (1) a forecasting dummy ($F=0$ for games having forecasting); (2) an information dummy ($C=0$ for complete information); (3) time (t = the index of the period, counting only those the player has been in whose price change is the dependent variable in the observation); (4) the lagged price of the player whose price change is the dependent variable of the current observation ($Plg$); (5) the difference betwen $Plg$ and the contemporaneous mean of the prices of the other players in the game ($Plg-\overline{P}$); (6) $\Delta Plg$ (the most recent nonzero changes in $Plg$); and (7) $\Delta\overline{P}$ (the most recent nonzero changes in $\overline{P}$). All results reported in this paper are for the two-firm symmetric game.

We used these regressions to define steps. A step is any price change which lies at least two standard deviations away from the regression line. We removed all steps from the samples and reran the regressions. These regressions are intended to explain price choice when price is changed in a routine way. The price change regressions give the following equations:[2]

$$\text{(5)}\quad \Delta P = .024 + .035^{3}F + .032C + .00022t - .179^{3}Plg - .503^{3}(Plg - \overline{P}) + .088^{3}\Delta Plg - .042\Delta\overline{P}$$

for decreases and

$$\text{(6)}\quad \Delta P = .161 - .0115F - .069^{3}C - .000094t - .072^{3}Plg - .319^{3}(Plg - \overline{P}) - .0022\Delta Plg - .022\Delta\overline{P}$$

for increases. The equations may be thought of as conditional reaction functions, conditional on the direction of price movement.

Returning to the identification of signals, we used two criteria to define signals. A signal should be a particular con-

[2] These regressions also had one dummy variable for each subject. These are not shown here. Equations (5) and (6) show intercept terms for one particular subject.

[3] These coefficients are significant at the 5 percent level.

TABLE 1

| | | $P-\overline{P}$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | −.25 | | | 0 | | | .25 | | |
| | .25 | .589 | .380 | .031 | .550 | .254 | .196 | .510 | .059 | .430 |
| $P$ | .50 | .635 | .341 | .025 | .597 | .239 | .164 | .559 | .059 | .382 |
| | 1.00 | .719 | .265 | .015 | .686 | .203 | .111 | .651 | .057 | .292 |

spicuous action which has almost no *direct* effect on the firm's profit for the game. Pulses meet these conditions because they are conspicuous and, by returning quickly to the original price, they cause little change in profits (directly). Steps, though conspicuous, appear to violate the second condition because they seem to make a long-term shift in the level of a firm's price.

### B. Price Change vs. No Price Change

To estimate the probability of price change, we used logit analysis.[4] This analysis takes into account the presence of pulses in the data. For an observation, $P_+$ is a dummy variable whose value is 1 if a rival firm has finished a pulse in one of the three immediately preceding periods. $P_+=0$ otherwise. $P_-$ is analogously defined for negative pulses. Letting

$$(7)\quad a = .236 - .324F - .197C - .0012t - .777^5 Plg + .634(Plg - \overline{P}) - .638^5 P_+ + .161P_-$$

the estimated probability of no change in price is $1/(1+e^a)$ for the two person symmetric game. Again using logit, with

$$(8)\quad b = .160 + .312F + .371C - .00006t + .457Plg - 8.973^5(Plg - \overline{P}) - 1.701P_+ + .533P_-$$

the estimated probability of a price increase, conditional on price change taking place, is $e^b/(1+e^b)$.

[4] See David Brownston et al. (1975) and Daniel McFadden (1974).

[5] Coefficients divided by their standard errors exceed 2.

To give an idea of what probabilities emerge, Table 1 shows the probability of no price change, price increase and price decrease for various values of $Plg$ and $Plg-\overline{P}$ with $t=200$, $F=0$, $C=0$, $P_+=0$ and $P_-=0$. The single period Cournot equilibrium occurs at $Plg=.5$ and $Plg-\overline{P}=0$. For each column and row there are three entries. The first is the probability of no change, the second, of increase and the last, of decrease.

The lower is a firm's price, the higher the probability it will be increased; and the lower it is in relation to the other prices, the higher the probability of increasing it.

### C. The Probability of Signals

Again, using logit, we estimated the probability of emission of a signal. As these are rare events, we failed in attempts to relate emission to signals received. With

$$(9)\quad c = -2.825^5 - .191F + .029C - .002t - .648Plg + 1.133(Plg - \overline{P})$$

the probability of a signal is $e^c/(1+e^c)$. Letting

$$(10)\quad d = .422 - .686F - .211C + .0008t - .759Plg - 7.222^5(Plg - \overline{P})$$

the conditional probability of emission of a positive pulse, given that a pulse is emitted, is $e^d/(1+e^d)$. Probabilities for no signal, plus signal and negative signal are given in Table 2. In Tables 2 and 3 the labeling of rows and columns is identical to Table 1. Both negative and positive signals are

TABLE 2

| | | $P-\overline{P}$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $-.25$ | | | 0 | | | .25 | | |
| | .25 | .975 | .022 | .002 | .967 | .019 | .013 | .957 | .008 | .034 |
| $P$ | .50 | .979 | .019 | .003 | .972 | .015 | .013 | .963 | .006 | .031 |
| | 1.00 | .985 | .013 | .003 | .980 | .009 | .011 | .973 | .003 | .024 |

more probable the lower is a firm's price. Plus signals are more probable as the firm's own price is lower in relation to the prices of the others, while the probability of negative signals rises as the firm's own price rises relative to the other's price.

### D. *Putting Pieces Together to Describe Expected Price Behavior*

The information collected may be thought of as a rather complicated reaction function. The regression equations for price increases and decreases are, by themselves, like ordinary reaction functions; however, they estimate price changes conditional upon change taking place. There is at any time a substantial probability that no change will occur. An equilibrium vector of prices in the present context is characterized by prices such that the expected price change is zero. This means that for each player the probability of price increase multiplied by the expected size of an increase (given that an increase will occur) plus the probability of decrease multiplied by the expected size of a decrease should be zero. For the two-person symmetric game, this occurs at a price of .826 if t is taken to be 200 and at .822 if t is 300. The Cournot equilibrium is at $P=.5$ and $P-\overline{P}=0$, while the joint profit maximum is at $P=1$ and $P-\overline{P}=0$. As signals occur rarely, they have small impact on equilibrium behavior. To examine their *local* impact we imagined, contrary to possibility, that signals could be sent in every period. This shifts the equilibrium above the joint maximum price for negative signalling and below the Cournot price for positive signalling. These are strong effects.

Table 3 shows the expected value of a price increase, given that increase will take place, the expected value of a decrease, given that a decrease occurs, and the expected value of price. For example, at the Cournot equilibrium ($p_i=.5$, $P_i-\overline{P}_i=0$), these are .125, $-.113$ and .011.

The prices at which expected price change is zero make sense in comparison with earlier experimental research. The general behavior described above is also plausible, though it differs in a conspicuous way from the theoretical literature on reaction functions. That literature sup-

TABLE 3

| | | $P-\overline{P}$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $-.25$ | | | 0 | | | .25 | | |
| | .25 | .223 | .057[a] | .087 | .143 | $-.069$ | .023 | .063 | $-.194$ | $-.080$ |
| $P$ | .50 | .205 | .012[a] | .070 | .125 | $-.113$ | .011 | .045 | $-.239$ | $-.089$ |
| | 1.00 | .169 | $-.077$ | .044 | .089 | $-.203$ | $-.004$ | .009 | $-.329$ | $-.096$ |

[a] These values should, of course, be negative. They result from extrapolation beyond the main range of the data.

poses that as soon as a steady state pair of prices is reached, there is no further price change; and before such prices are reached, prices change in each period. Our subjects behave differently. The counterpart of the steady state is where expected price change is zero; however, we would not see zero change here. Meanwhile subjects do not act on price most of the time, so, even away from the "equilibrium" no price change is more probable than change. Furthermore, prices are not changed on a regular schedule. Whether these differences are important or only cause minor random deviations remains to be seen.

A difference of clear substance is the divergence between price increases and price decreases, and the presence of signals. The more formal reaction function literature formulates present price as a function of the past prices observed in the market. There is no asymmetry of behavior between increases and decreases. By contrast, our experimental evidence suggests that subjects make a three-way decision for routine price behavior where the alternatives are increase, decrease and no change. Each of the first two has its own rule (equation) to determine size. Superimposed on this is signaling where unusual price movements are made to apparently convey a message. We have not gotten much beyond identifying these. It would take further study and a very large sample to see their role clearly.

REFERENCES

**D. Brownston, G. M. Duncan, and D. McFadden,** *Quail User's Manual,* working pap. no. 7402; Institute of Transportation and Traffic Engineering, Univ. of California, Berkeley, Jan. 1975.

**L. E. Fouraker and S. Siegel,** *Bargaining Behavior,* New York 1963.

**J. W. Friedman,** "An Experimental Study of Cooperative Duopoly," *Econometrica,* 1967, *35,* 379–97.

**A. C. Hoggatt,** "Response of Paid Student Subjects to Differential Behaviour of Robots in Bifurcated Duopoly Games," *Rev. Econ. Studies,* 1969, *36,* 417–32.

**D. McFadden,** "Conditional Logit Analysis of Qualitative Choice Behavior," Ch. 4 in *Frontiers of Econometrics* (Paul Zarembka, ed.), New York 1974.

# An Experimental Market for Public Goods: The *PBS* Station Program Cooperative

*By* John A. Ferejohn and Roger G. Noll*

Beginning with the 1974–75 season, the Public Broadcasting Service (*PBS*) undertook a three-year experiment to decentralize through a market process the selection of programs to be broadcast over the national noncommercial television network. The experimental market, called the Station Program Cooperative (*SPC*), enables each of the approximately 150 participating stations to "vote with dollars" for the programs they prefer, based on information about the subject matter, quality and costs of the alternatives.

The *SPC* had two purposes. The first was to determine whether a decentralized market mechanism for acquiring national programming can be established that, first, has a reasonable cost and, second, retains the basic features of a network. A network is a means of centralizing the acquisition, transmission and promotion of programming. Although the advantages of networking are several, probably the most important is that it saves substantial transactions cost in comparison to the two other alternatives: syndication (in which each station negotiates for nationally distributed programs with the owner of the rights to the program) and localism (in which each station produces its own programming). (See Bruce Owen, Jack Beebe and Willard Manning.)

The second purpose of the *SPC* is to determine its impact upon the types of programs that are broadcast. Centralized programming authorities such as the present commercial networks and the Corporation for Public Broadcasting, because they are few in number, are relatively easy targets for political pressure with respect to program content, may at least tacitly collude in programming decisions, and in any event, have limited experience concerning the nature of the demand for programs in all markets in the network system. (See Noll, Merton Peck and John McGowan, Ch. 8.) Because of the cost advantages of networking and the technical barriers to entry of new networks, in principle a network can produce an inefficient programming mix and still meet minimum standards of profitability or, in the case of the noncommercial sector, public acceptability. Concomitantly, station managers may be too unsophisticated about judging programming quality and sensing the structure of demand in the local market to select programming that, in benefit-cost terms, is "better" than that selected by a centralized decision maker. While direct evidence on which system works better is not possible to acquire, two useful questions can be answered, at least in principle: (1) Do the *SPC* and centralized decision making lead to the selection of different programs, and if so, is that difference systemmatic? and (2) Which system produces higher audience

* California Institute of Technology. Financial support for preparing this paper was provided by the National Science Foundation RANN program, grant #APR76-01920. We are indebted to Richard Beatty and *PBS* for providing us with remarkably complete data for all three *SPCs*. We are, of course, solely responsible for the contents of this paper.

ratings and/or greater total revenues (private contributions plus government appropriations) for the noncommercial television network? This paper does not address these issues. (See Natan Katzman and Noll.) That research is still underway.

Though unintentional, the *SPC* experiment has a third, more general use. The *SPC* is a test of a novel mechanism for public decision making about a particular class of public goods, specifically those for which binary (all or nothing) exclusion is possible.

The purpose of this paper is to make the connection between the *SPC* and the theory of public goods, and to draw some preliminary conclusions about the workability of this kind of artificial market. It is, therefore, addressed to the first and third issues.

A television broadcast is a public good in that the cost of transmitting the program within a given geographical area is independent of the number of television sets that are tuned to it. So, too, is the program a public good as far as each station is concerned. Excluding rents arising from differing popularities of programs, the costs of a program are independent of the number of stations over which it is broadcast. Even national network distribution costs are, in most instances, independent of the number of stations receiving the program. Nearly all programs distributed by networks are carried over the national telecommunications system. If the program is distributed from coast to coast, the transmission costs are only slightly dependent on the number of intermediate points at which the program is received and broadcast. As the distribution of television programs moves to satellite transmission, national interconnection costs will become totally independent of the number of stations in the network.

Nevertheless, exclusion of stations from use of a particular program is also essentially costless. In both the commercial and noncommercial systems, few stations accept ("clear") all network programs. One way to view the process by which stations make programming decisions is that each station has a preference ordering over all available programming possibilities and selects the most preferred feasible set of programs, given its budget and the costs of the alternatives. Since programs are public goods, these decisions have external effects; namely, if a program is selected by one more station, costs per station decline, and hence from the vantage point of other stations some programming alternatives involving this program that dominate a previously selected alternative may become feasible.

## I. The Mechanics of the *SPC*

The heart of the *SPC* is a mechanism for calculating the price of each program for each station that broadcasts it. These prices are determined through a sequential process in which a series of trial price vectors are proposed, responded to, checked for acceptability with program producers, and, if they generate too little or too much revenue, updated.

The first stage of the *SPC* is the dissemination of program proposals. Program producers—primarily stations, but also some independents such as the Childrens Television Workshop ("Sesame Street")—provide descriptions of proposed programs (series or special events), prices, and in many cases pilots. The form of the final proposal is usually negotiated with *PBS*, which among other things, tries to keep program prices roughly in line with production costs. In each of the first three years of the *SPC*, about 250 programs were proposed.

The second stage of the *SPC* culls the list of proposed programs in the following manner. On the basis of the final pro-

posals, stations are asked to assign "priorities" to each program, which amounts to rating the programs on a scale of one to five (five is best). The sum of the ratings for each program is calculated, and the programs are rank-ordered according to their scores. Programs ranking in the top third, plus a few others which *PBS* believes deserve special consideration (their classifications show high variance, or additional information about the program will become available and might change scores), are then allowed to enter the market phase of the selection process. This process served to cull the list of contenders to 93 in the first year (*SPC* I), and 136 in the second (*SPC* II).

Next, the market phase of the *SPC* is begun. Its main feature is an algorithm for calculating the program prices facing each station. The formula for calculating the price $P_{ij}$ of program $j$ to station $i$ is:

$$P_{ij} = C_j \left( .8 \frac{B_i}{\sum_{k \in Sj} B_k} + .2 \frac{N_i}{\sum_{k \in Sj} N_k} \right)$$

where

$C_j$ = the cost of program $j$
$B_k$ = the programming budget of station $k$
$N_k$ = the population of the area served by station $k$
$S_j$ = the set of indexes of stations that have agreed to purchase program $j$.

Obviously, prices and the identity of purchasing stations are interdependent. In the *SPC*, the initial price calculations are based upon the assumption that each program is selected by stations representing 80 percent of the combined budget and total audience of the entire public television system. In subsequent rounds, prices are calculated on the basis of the budgets and populations of the stations selecting the program in the previous round. Once the market has eliminated all but a few more programs than are likely to be purchased, the *SPC* manager announces that in subsequent rounds stations will be required to continue selecting a program that they selected in the previous round if the price does not increase. The market process continues until all programs either generate revenues acceptable to the producer of the program, receive no votes from other than the producing stations at prices based upon selections made in the previous round, or are withdrawn from consideration by the producer.[1] Then a final "late purchase" round takes place in which each station can select additional programs from the list of purchased programs. Stations eventually purchased 25 programs in *SPC* I and 38 in *SPC* II, each station buying about 20 in the first and 30 in the second.

The costs of programs are only upper bounds on the amount of revenue producers eventually receive from the *SPC*. At any phase of the market process, a producer may guarantee that the current price vector for his program will not rise —that is, that the producer will absorb the shortfall in revenues in relation to the listed total cost. Since many of the programs entered in the *SPC* will be produced in any event for local or regional audiences, any revenue the producer receives from the national market is a net gain. Similarly, coalitions of buyers are permitted to guarantee the stability of current prices by agreeing to make up the difference between revenues and costs.

## II. The Economics of the *SPC*

The appropriate theoretical framework for analyzing the properties of the *SPC* is the extension of general equilibrium

[1] The purchase rounds are conducted through a nationwide computer network. Each station manager sits at his teletypewriter, receives messages about the progress of the *SPC* (including price updates), and each round enters his selections.

theory that includes pricing of public goods. Models of optimal choice of political jurisdiction, the theory of clubs without congestion, and Lindahl pricing of public goods are obvious examples. (See James Buchanan, Bryan Ellickson, Duncan Foley, Eric Lindahl, R. W. Rosenthal and Paul A. Samuelson). In a market with $p$ programs, a particular pattern of program selection and rejection can be represented by a $p$-dimensional vector in which the elements are either zero or one according to whether a program was rejected or selected by the station. A family of $2^p$ such vectors describes all possible decisions by a station, and presumably a station is able to generate a transitive weak ordering over these vectors that reflects its programming preferences. The problem faced by a station is to pick the highest-ranking selection vector—e.g., a vertex on the $n$-dimensional unit square—that is feasible, given its budget and the prices of the programs.

Each program producer seeks to maximize the revenue it derives from a program, subject to two constraints: the revenue received through the *SPC* must cover the incremental cost of allowing the program to become national, and it must not exceed the maximum allowable costs negotiated with *PBS*, which are intended to be approximately total production costs. The minimum and maximum prices are identical for programs that will be produced only for national distribution unless the producer can find other sources of financial support, is willing to take a loss, or is a station that values the program idea sufficiently highly that it is willing to pay more than its share of the costs to guarantee its production.

The *SPC* is a process of generating for each station a price vector and a vector of selections and for programs a cost vector that constitute a stable equilibrium in the feasible production and budget sets. This algorithm has two features: an initialization procedure that generates the first trial set of prices and an updating procedure that changes prices until a stable equilibrium is found.

From a theoretical standpoint, it is a miracle that the process works. One source of difficulty is that the production set is not convex. The space in which programs are located is of an unknown number of dimensions of program characteristics. Past research has been notably unsuccessful in defining these dimensions in an empirically useful way (see Edward Greenberg and Harold Barnett), but as a working hypothesis a program can be thought of as a particular combination of production talents, production quality (e.g., the average number of "takes" per printed scene, the number of scenes filmed on location, etc.), and subject matter. Each program proposal in the *SPC* is, to all practical purposes, a single point in this space—not a ray, but a point—because essentially no variance is possible during the operation of the *SPC* in any of the relevant dimensions, including the number of units of the program to be sold. As a result, linear combinations of programs cannot be formed to arrive at locations in program-characteristics space that are between the program points. Consequently, even if at all factor prices one unit of a particular program would never be selected in market competition with one unit of each of the other programs, it may still be selected in the *SPC* since the combinations of other programs that strictly Pareto dominate it cannot generally be formed.

The first two *SPCs* had another potential problem arising from the initialization procedure. Because the initial price vector was calculated from grossly unrealistic assumptions about the number of stations that would select each program, the station selection vectors they called forth were

necessarily outside the system's feasible set. In *SPC* I second round prices averaged nearly four times first-round prices! This kind of initialization causes a massive reduction in the number of program selections per station in the first few rounds. If initial selection vectors are sufficiently diverse that few programs appear clearly out of the running in the initial stages, stations will not quickly concentrate votes on a relatively few popular offerings, and so for nearly all programs prices will continue to rise through the first few rounds. If preferences are sufficiently diverse and the number of programs large enough compared to the number of stations, the process might not converge to the feasible set. One can imagine a desperate round in which each station, facing astronomical prices, picks exactly one program, all stations pick a different program and no program can cover costs on the budget of a single station.

Theoretically, whether initialization and updating procedures for the *SPC* can be derived that reach a stable equilibrium is as yet unanswered. Our conjecture is that they can, particularly if the procedures constitute an improvement algorithm rather than a search for the feasible set from outside it.

The third *SPC* will try a new initialization procedure. Instead of calling out a selection vector and observing a budget-breaking consequence, each station will provide the *SPC* with a total expenditure limit and a preference ordering of the programs. The *SPC* will then calculate new prices based on these preferences and tell each station how far down its preference ordering it can go, given its budget constraint. For the first round, preference ordering will not be based on an explicit price formula as stations will draw their own inferences from program cost information. In subsequent rounds, the same price updating procedure used in the other *SPCs* will be used. The basic idea is first to find a point on the boundary of the feasible set, and then move along the boundary until an equilibrium is reached. The main difficulty with the procedure is finding a nontrivial initial feasible solution.

### III. The Results of the First Two *SPCs*

The *SPC* results of interest here relate to the convergence properties of the system; we will not investigate the issues related to the types of programs the system selected.

In both of the first two *SPC* experiments, the market phase of the selection process was allowed to run for twelve rounds. *SPC* I had two-thirds as many programs from which to choose, and eventually purchased two-thirds as many programs, as *SPC* II.

By round 6 in *SPC* I and round 8 in *SPC* II, the programs that were purchased had all but been determined. In *SPC* I, the three programs still in the market during round 6 that were subsequently eliminated received support from 3, 5 and 8 stations, respectively, while the three lowest numbers of supporting stations for programs that were eventually purchased were 19, 21 and 34. Nineteen of the 25 programs that were eventually purchased were, by round 6, receiving support from over half of the stations (75).

In round 8 of *SPC* II, the six programs that were eventually dropped received support from 2, 9, 36, 40, 64 and 66 stations, respectively. Of the 38 programs that were eventually purchased, 31 were supported by more than half the stations in round 8. Five more received support from between 58 and 73 stations and two that were eventually purchased received only 40 and 42 votes in round 8. One of the latter was clearly a special case, consisting of 100 sixty-second messages promoting coming attractions. Thus, of

the ten conventional programs receiving support from between 36 and 73 stations, six were purchased. All six that were purchased had total costs below $15,000 per program-hour, which is inexpensive. Of the four that were rejected, hourly cost data are available only for two. One was a children's program that cost $57,000 per hour—more than triple the cost of "Sesame Street" or "The Electric Company"—and the other was a documentary series that cost more than $100,000 per hour. Thus, further classification of programs by cost apparently cleanly separates the purchased from the excluded programs.

In both markets, the first few rounds saw massive reductions in the number of programs selected per station and generally rising program prices. In *SPC* I, prices for programs remaining in the market rose abruptly at the start, and did not, on average, begin to fall until after the seventh round, while in *SPC* II, prices began to fall after round 3.[2] In both markets, the average price of purchased programs did not return to the average round 1 initialization price until the last round.

The first two *SPCs* appear to have reached an equilibrium, at least within the context of the operating rules. Although the twelve purchase rounds occurred over only a few days, late purchases were possible for several weeks. Two programs in *SPC* I and seven in *SPC* II increased the number of stations supporting them by 3 to 8 percent during the late purchases. In both markets about 55 percent of the programs made no additional sales during the late purchase, while 28 percent in *SPC* I and 26 percent in *SPC* II increased their number of supporters by less than 2 percent.

Outside the context of the *SPC* rules, the final list of programs purchased may not be an equilibrium. The key operating rule in the *SPC* appears to be the non-reincarnation rule: once eliminated, a program cannot reenter the market at a later stage. This eliminates the possibility of cyclic selections, but may also prevent the system from reaching an efficient equilibrium if one exists. About one-third of the programs that eventually were purchased received twice or more as many votes in the final selection round as they did at their ebb during the market. Many of these programs also had mediocre pre-market rating scores, averaging below a ranking of three in the five point scale.

With the exception of buyer and seller guarantees as explained above, stations did not have an opportunity to price-discriminate against themselves in order to register an intense preference for a particular program. A group of stations with an intense preference for a particular program, A, but unable to guarantee its purchase, could not offer remaining stations a discounted price to tempt them away from another program, B. If the remaining stations had only a slight preference for B over A, the pricing rule could prevent a substitution of A for B at prices that would make the move a Pareto improvement.

## IV. Conclusions

Because our research on the *SPC* has only begun, few conclusions can be drawn as yet. Most of the inferences we are prepared to draw relate to the additional research that is required to gain a better understanding of how the system operates, if it could be adapted to the acquisition of other public goods (national parks, commercial broadcasting, etc.), and whether it produces acceptable programming results at reasonable transactions cost.

[2] This statement is true for both the unweighted average price and the average of prices weighted by the number of stations selecting the associated program. Movements of the latter index the change in the average budget of a station, holding its selections at the mean of all stations.

An obvious candidate for further work is theoretical examination of the equilibrium properties of the *SPC*. Two types of questions are relevant: What normative properties, if any, do the existing *SPC* rules have, and would other initialization and updating procedures work better? Although of some interest is whether the system reaches an equilibrium that is Pareto optimal, for policy purposes the conclusion that it does not is not persuasive that the *SPC* should be abandoned. A centralized system, for reasons outlined briefly above, also is likely to be inefficient. Of prime practical interest are the consequences of alternative procedures for operating the system.

Laboratory experiments could provide a useful additional source of information about alternative structures of *SPC*. Very briefly, the idea of each experiment would be to ask a group of subjects to select an unstated number of items from a list of alternatives. Total costs for each alternative would be provided, as well as a payoff function for each group member. An initialization and updating procedure would be adopted, and the general mechanics of the *SPC* followed. By running the identical experiment several times, the uniqueness of the results of the system could be checked. By varying payoff functions while holding procedures fixed, the sensitivity of the system to intense preferences could be explored. And by varying procedures while holding payoffs constant, the performance of various initialization and updating methods—particularly of improvement algorithms and of other procedures—could be compared. The advantage of the experimental system, of course, is that it is far less expensive—in terms of operating costs, risks to something of significant value (the public television system), and the time between observations—than, through the years, manipulating the *SPC*.

Finally, the overriding conclusion of our investigations to date is that experimental markets for public goods are an interesting, potentially useful and almost completely unexplored focus for research. If this paper conveys to other social scientists our perception of the fascinating possibilities suggested by the *SPC*, its most important purpose will have been served.

## REFERENCES

J. M. **Buchanan**, "An Economic Theory of Clubs," *Economica*, Feb. 1965, *32*, 1–14.

B. **Ellickson**, "A Generalization of the Pure Theory of Public Goods," *Amer. Econ. Rev.*, June 1973, *62*, 417–32.

D. **Foley**, "Lindahl's Solution and the Core of an Economy with Public Goods," *Econometrica*, Jan. 1970, *38*, 66–72.

E. **Greenberg and H. J. Barnett**, "TV Program Diversity—New Evidence and Old Theories," *Amer. Econ. Rev. Proc.*, May 1971, *61*, 89–93.

N. **Katzman**, *Public Broadcasting's Station Program Cooperative*, Aspen Program on Communications and Society, Palo Alto 1974.

E. **Lindahl**, "Just Taxation—A Positive Solution," in *Classics in the Theory of Public Finance* (R. A. Musgrave and A. T. Peacock, eds.), St. Martin 1964.

R. G. **Noll**, "Decentralization of Public Television," in *Conference on Communications on Communications Policy Research: Papers and Proceedings*, Office of Telecommunications Policy, Washington 1973.

———, **M. J. Peck and J. J. McGowan**, *Economic Aspects of Television Regulation*, Brookings Institution, Washington 1973.

B. M. **Owen, J. H. Beebe and W. G. Manning, Jr.**, *Television Economics*, Lexington 1974.

R. W. **Rosenthal**, "External Economies and Cores," *J. Econ. Theory*, June 1971, *3*, 182–87.

P. A. **Samuelson**, "Pure Theory of Public Expenditure and Taxation," in *Public Economics* (J. Margolis and H. Guiton, eds.), St. Martin 1969.

# Experimental Economics: Induced Value Theory

By VERNON L. SMITH*

It is the premise of this paper that the study of the decision behavior of suitably motivated individuals and groups in laboratory or other socially isolated settings such as hospitals (R. Battalio, J. Kagel, et al., 1973) has important and significant application to the development and verification of theories of the economic system at large. There are two reasons for this.

1. The results of laboratory studies can serve as a rigorous empirical pretest of economic theory prior to the use of field data tests. The state of economic hypothesis testing, as it is sometimes done, can be described roughly as follows: based on casual observation of an economic process and the self-interest postulate, one develops a model, which is then tested with the only body of field data that exists. The results of the test turn out to be ambiguous or call for improvements, and one is tempted to now modify the model in ways suggested by the data "to improve the fit." Any test of significance now becomes hopelessly confused if one attempts to apply it to the same data. Where it is possible and feasible, as in the study of price formation, the data from controlled experiments can be used to test hypotheses stemming from prescientific casual observations of a particular phenomenon. The fact that one can always run a new experiment means that it is never tautological to modify the model in ways suggested by the results of the last experiment. Since economic theories always deal with certain alleged behavioral tendencies in isolation, the experimental laboratory is uniquely well suited for testing the validity of such theories. It provides an exceptionally rigorous discipline of our ability to model elementary situations whether or not field data can be regarded ultimately as having been generated by such elementary models.

2. The results of experiments can be directly relevant to the study and interpretation of field data. Other so-called nonexperimental sciences such as meteorology and astronomy have depended crucially for their development on (a) small-scale laboratory experiments in the physics of mass motion, thermodynamics, and nuclear reactions; and (b) the postulate that such microphysical experimental results apply, with suitable modifications, to the study of the weather, the planets and the stars. This parallelism, "As far as we can tell, the same physical laws prevail everywhere" (Harlow Shapley 1964, p. 43), also has application to the study of social economy. Laboratory experience suggests that all of the characteristics of "real world" behavior that we consider to be of primitive importance—such as self-interest motivation, interdependent tastes, risk aversion, subjective transactions cost (time is consumed), costly information (it takes time to acquire and process information), and so on—arise naturally, indeed inevitably, in experimental settings. Anyone who had begun the study of economics in

* Department of Economics, University of Arizona. Support from NSF grants is gratefully acknowledged. This paper is an articulation of concepts originally developed in the course of several seminars in experimental economics taught at Purdue University, 1964–67.

the laboratory without these concepts would soon find himself inventing them. Furthermore, the process of experimental design forces one to articulate rules and procedures, the collection of which forms an institution, organization, or "body of law" with striking "real world" parallels (cf. Martin Shubik 1974). The laboratory becomes a place where real people earn real money for making real decisions about abstract claims that are just as "real" as a share of General Motors.

## I. The Theory of Induced Valuation

Control is the essence of experimental methodology, and in experimental exchange studies it is important that one be able to state that, as between two experiments, individual values (e.g., demand or supply) either do or do not differ in a specified way. Such control can be achieved by using a reward structure to induce prescribed monetary value on actions. The concept of induced valuation (Smith 1973) depends upon the postulate of *nonsatiation:*

> Given a *costless* choice between two alternatives, identical except that the first yields more of the reward medium (usually currency) than the second, the first will always be chosen (preferred) over the second, by an *autonomous* individual, i.e., utility is a monotone increasing function of the monetary reward, $U(M)$, $U' > 0$. [pp. 22–23]

This postulate applies to experiments designed to test price theory propositions conditional upon known valuations. Separate experiments can be designed to test propositions in preference theory.

Example 1. In the experimental study of competitive equilibria in isolated markets it is necessary to induce known (to the experimenter) supply or demand on individual subjects. Let subject buyers $i=1, 2, \ldots, n$ each be given a table listing increasing concave total receipts $R_i(q_i)$ representing the currency redemption or "resale" value of $q_i$ units acquired by subject $i$ in an experimental market. The instructions state that if subject $i$ acquires $q_i$ units at prices $p_1^i, p_2^i, \ldots, p_{q_i}^i$, he will receive cash earnings of $R_i(q_i) - \sum_{k=1}^{q_i} p_k^i$. Neoclassical demand is defined as the quantity that would be purchased as a function of a given hypothetical price $p$. By this definition, if for a fixed $p$ a subject purchases $q_i$ units, he earns $R_i(q_i) - pq_i$. If his utility for money is $U_i(M_i)$ he will wish to $\max_{q_i} U_i[R_i(q_i) - pq_i]$. We have an interior maximum if and only if

$$(R_i' - p)U_i' = 0,\ U_i' > 0, \qquad \text{or } q_i = R_i'^{(-1)}(p),$$

for the class of functions $U_i$, $R_i$ such that $(R_i' - p)^2 U_i'' + U_i' R_i'' < 0$. This reward scheme induces arbitrary demand $R_i'^{(-1)}(p)$ on subject $i$, and the experimentally controlled market demand becomes $Q = \sum_{i=1}^{n} R_i'^{(-1)}(p)$ independent of the $U_i$.

Similarly, let $j=1, 2, \ldots, m$ subject sellers be given cost functions $C_j(q_j)$, and receive cash earnings $\sum_{k=1}^{q_j} p_k^j - C_j(q_j)$ from selling $q_j$ units at prices $p_1^j, p_2^j, \ldots, p_{q_j}^j$. If utility is $V_j(M_j)$, $V_j' > 0$, then $\max_{q_j} V_j[pq_j - C_j(q_j)]$ implies a supply function $q_j = C_j'^{(-1)}(p)$. The experimentally controlled market supply is $Q = \sum_{j=1}^{m} C_j'^{(-1)}(p)$ independent of the $V_j$. Such induced supply and demand become flows per period in experiments in which trading is conducted in a sequence of periods.

Example 2: Let subject traders be given a table listing increasing concave currency receipts $M(x_1, x_2)$ to be paid by the experimenter for terminal stocks $(x_1, x_2)$ of each of two abstract experimental commodities exchanged in an experimental general equilibrium market. Then subject $i$'s unknown utility for currency $U_i(M)$ induces the value $U_i[M(x_1, x_2)]$ on terminal stocks $(x_1, x_2)$. Consequently, the experimentally controlled indifference map given by the level contours of $M(x_1, x_2)$ are induced upon subject $i$ independent of his particular $U_i$. That is, each subject's marginal rate

of substitution of $x_2$ for $x_1$ is given by $U_i' M_1/U_i' M_2 = M_1/M_2$, $U_i' > 0$. This allows the "Edgeworth Box" representation of general exchange equilibrium to be reproduced experimentally by inducing a given indifference map on each member of one group of subjects, and another indifference map on each of a second group of subjects. With given endowments of the abstract commodities for members of each of the two trading groups, the experimental stage is set for exchange.

## II. Some Qualifications

There are three important qualifications to the nonsatiation postulate:

1. There may be subjective costs (or values) associated with market decisions. In a competitive market experiment a subject may find it arduous to monitor and make quotations, and to execute transactions. If such considerations are not negligible, then we lose some control over the process of induced valuation. The effect of boredom and the subjective costs of decision making have been emphasized in the important study by Sidney Siegel (1961). Roger Sherman (1974) has interpreted alleged violations of the Savage axioms in terms of the subjective cost of making the appropriate computations. In terms of the utility interpretation of the previous section, the utility function can now be written $U^i(M_i, E_i)$ where $E_i$ is the "transactional effort" required to obtain reward $M_i$ (cf. Harvey Leibenstein 1969; and implicitly, Ronald Coase 1960). To see the potential implications of costly choice, consider example 1 of the previous section in which demand $R_i'^{(-1)}(p)$ is induced upon $i$. Utility is now $U^i\{R_i[q_i(E_i)] - pq_i(E_i), E_i\}$ where it is assumed crudely that "bargaining effort," $E_i$, results in the purchase quantity $q_i(E_i)$. Then $\max_{E_i} U^i$ implies $(R_i' - p)\, q_i' U_1^i + U_2^i = 0$, and now the induced demand is $q_i = R_i'^{(-1)}(p - U_2^i/U_1^i q_i') < R_i'^{(-1)}(p)$, if $U_2^i < 0$, $q_i' > 0$. Hence, if there is a cost (value) to transacting in the experimental task, the induced demand will be smaller (larger).

There are several ways of dealing with this problem:

(a) One is to examine the experimental results to see if the quantity exchanged is less than predicted. If it is, this is consistent with a significant transactions cost. Awareness of such transactions cost may provide valuable clues to understanding why certain experiments may fail to produce predicted results. The process is not tautological as long as one can redesign the experiment and show that such conjectured transactional effects can be reduced.

(b) Another approach is to use a reward structure to compensate for, or offset, the subjective costs of transacting. There are two ways of doing this. (i) One way (Siegel 1961) is to simply raise the reward level. This increases the subjective value relative to the subjective cost of acquiring units $q_i$. Let $\alpha$ be a scale parameter determining reward level. Then utility becomes $U^i\{\alpha(R_i[q_i(E_i)] - pq_i(E_i)), E_i\}$. Induced demand is now $q_i = R_i'^{(-1)}(p - U_2^i/U_1^i q_i'\alpha) \rightarrow R_i'^{(-1)}(p)$ in the limit as $\alpha$ increases provided that the marginal rate of substitution $-U_2^i/U_1^i q_i'\alpha$ decreases with the reward level. (ii) Alternatively, and this is the device used most extensively, subjects are promised a "commission," $\beta$, for *each* transaction in addition to their cash trading profits. Now utility is $U^i\{R_i[q_i(E_i)] - (p-\beta)q_i(E_i), E_i\}$, and induced demand is

$$
\begin{aligned}
q_i &= R_i'^{(-1)}(p - \beta - U_2^i/U_1^i q_i') \\
&\cong R_i'^{(-1)}(p) \qquad \text{if } \beta \cong -U_2^i/U_1^i q_i' > 0.
\end{aligned}
$$

Compare two experiments (Charles Plott and Smith 1975, pp. 20–21) in which the induced supply and demand conditions were identical but one paid no cash trading commission, only trading profit, while the other paid both: In the one experiment,

volume was below (17–18 units) the "theoretical" equilibrium quantity (20 units) in all seven trading periods; in the second experiment, volume was below (19 units) equilibrium in only two of eight trading periods.

2. Individuals may attach game value to experimental outcomes. A profit in "points," $R_i(q_i)-pq_i$, may have subjective value $S_i[R_i(q_i)-pq_i]$. If $S_i$ is monotone increasing then such game utilities create no methodological problems since they reinforce rather than distort the effect of an explicit monetary reward structure. Because of such game utilities it is often possible in simple-task experiments to get satisfactory results without monetary rewards by using instructions to induce value by role-playing behavior (i.e., "think of yourself as making a profit of such and such when . . . "). But such game values are likely to be weak, erratic, and easily dominated by transactions costs, and subjects may be readily satiated with "point" profits.

Qualifications 1 and 2 are illustrated in the convergence behavior of three experimental markets with no cash rewards and seven markets with complete and with random cash rewards. In the first three cases subjects were asked to imagine that trading profits and commissions were real. In each case the market was organized as a continuous double auction. (Buyers could make oral bids and sellers oral offers for a single unit, and any seller could accept a bid, any buyer an offer. Each subject knew only his own demand or supply conditions.) (See Smith 1964, pp. 199–201 for the instructions.) In the first case (Smith 1962, p. 118, Chart 3) subjects trade only one unit per trading period. The absence of cash rewards does not hinder convergence to prices near equilibrium by the third trading period. However, deviations increase in period 4. In the absence of cash rewards this is more likely to occur as gaming boredom follows an initial (pleasant) experience of learning.

In a second experiment (previously unpublished) buyers received multiunit revenue (or resale value) schedules, and sellers multiunit total cost schedules. There were three buyers with one schedule, eight with another; four sellers with one cost schedule, eight with another. Now the task is more difficult and incentives are weak. Price convergence is strong, especially in the second period, since the greater volume when traders are given multiple-unit capacities increases the learning experience within a trading period. But volume is considerably below (24 and 26 units in the first and second periods) the competitive prediction (30 units). This is consistent with the above theory where the task is more difficult (higher transactions cost) and monetary rewards are absent.

Case 3 (Smith 1962, p. 119, Chart 4) illustrates an experiment which fails to reach either the competitive price or quantity although the market stabilizes nicely. In this case equilibrium requires contract prices to fall to the common limit price of all sellers. They are to "imagine" themselves as making a 5-cent commission on

TABLE 1—MEAN CONTRACT PRICE BY TRADING PERIOD

| Experiment | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Excess Supply | 5 | 5 | 5 | 5 | 8 | 8 | 8 |
| Reward Condition | Complete | Complete | Random | Complete | Complete | Complete | Complete |
| Information Condition | Incomplete | Incomplete | Incomplete | Complete | Incomplete | Incomplete | Complete |
| Trading Period 1 | 3.48 | 3.67 | 3.60 | 3.51 | 3.26 | 3.49 | 3.56 |
| Trading Period 2 | 3.29 | 3.26 | 3.44 | 3.40 | 3.15 | 3.28 | 3.25 |
| Trading Period 3 | 3.19 | 3.12 | 3.31 | 3.34 | 3.11 | 3.13 | 3.20 |
| Trading Period 4 | 3.14 | 3.10 | 3.24 | 3.37 | 3.10 | 3.12 | 3.17 |

trades at these limit prices, but clearly this is not real enough to induce many contracts at \$3.10 (the theoretical equilibrium). Not even a decrease in demand succeeded in lowering contracts to \$3.10 (Table 1). This contrasts with several experiments (1, 2, 5, 6 in Table 1) using complete cash rewards in which the supply and demand are even more asymmetric than in case 3. In Table 1, markets with an excess supply of five (eight) consisted of eleven buyers with limit prices \$4.20 and sixteen (nineteen) sellers with limit prices \$3.10. A different subject group participated in each double auction experiment. Convergence to the competitive price and quantity by trading period 4 was strong, although at the equilibrium price each buyer receives \$1.15 profit with commission per trade while each seller receives only the 5-cent commission.

A controlled measurement of the effect of complete versus random monetary rewards is shown in Table 1, experiments 1–3. In 1 and 2 all subjects were paid their trading profit plus commission in cash, while in 3 four of the 27 subjects were chosen at random to receive cash profits at the end of each trading period. The weaker random reward structure significantly retards the market's convergence.

Qualifications 1 and 2 lead to a precautionary corollary: with or without monetary rewards, the experimenter may be tempted to add "realism" by giving the abstract experimental commodity a name such as "wheat," or otherwise attempt to use instructions to simulate the alleged circumstances of a particular market. This runs the danger of so enriching induced values that control over valuation is lost. Suppose, as above, that a subject is paid $R_i(q_i)-pq_i$, but also perceives that he must attach instruction-induced value to $q_i$. Utility may now be $U^i[R_i(q_i)-pq_i,\ q_i]$, and demand becomes $q_i=R_i'^{(-1)}(p-U_2^i/U_1^i)>R_i'^{(-1)}(p)$. Consequently, it may be preferable *not* to embellish the instructions with well-intentioned attempts at "realism." Let the explicit reward structure be the singular source of valuation, insofar as this is possible.

3. Individuals may not be autonomous own-reward maximizers. Interpersonal utility criteria may qualify the theory of induced valuation. Thus subject $i$'s utility may depend upon both $i$'s and $k$'s reward, $U^i[R_i(q_i)-pq_i,\ R_k(q_k)-pq_k]$. If this condition prevails, then the demand of $i$ may depend upon that of $k$. However, this kind of interdependence is effectively controlled by the experimental condition of "incomplete" information, first defined and studied by Lawrence Fouraker and Siegel (1960, 1963) in experimental studies of bilateral bargaining and oligopoly. Under incomplete information subjects only know their own payoff contingencies. With $R_k(q_k)$ unknown to $i$, it cannot appear as a subjective argument of $U^i$.

The effect when subjects have complete information on each other's payoff contingencies is seen (Table 1) by comparing 1 (5) and 2 (6) with 4 (7). In 1 (5) and 2 (6) each subject knew only his own limit price. In 4 (7) the only change in the instructions was to add the information that there were eleven buyers, each with a \$4.20 resale value, and sixteen (nineteen in 7) sellers, each with unit cost \$3.10. From the mean price series it is seen that "complete" information of this kind retards the equilibrium tendencies of the double auction. Mean prices, especially in periods 3 and 4, tended to be higher under complete information than under incomplete information. The explanation is that with information on each other's payoffs, the way is open for "equity" considerations to modify self-interest choices. Sellers, believing that it is "fair" for trading profits to be shared between buyers and sellers, try to resist price decreases more vigorously than when they do not know what constitutes

such a fair price. Buyers acquiesce in this sharing by accepting many contracts well above $3.10, but since there is an excess of sellers, those holding out for the higher prices are the sellers most likely to fail to make contracts. Consequently, contract prices tend to decline, if slowly, when excess supply is 5, but more rapidly when excess supply is 8. The tendency of prices to be higher under complete information is contrary to the view of those who have argued that "perfect" information is essential for establishing competitive prices. The results are consistent with the game-theoretic proposition that more information increases the prospect of collusion (Shubik 1959, p. 171), and with the results of Fouraker and Siegel (1963, p. 187) in which the tendency of the competitive equilibrium to prevail under duopoly bargaining is reduced under complete information.

## REFERENCES

**R. Battalio, J. Kagel, J. Winkler, R. Fisher, R. Basmann and L. Krasner**, "A Test of Consumer Demand Theory Using Observations of Individual Consumer Purchases," *West. Econ. J.*, Dec. 1973, 411–28.

**R. Coase**, "The Problem of Social Cost," *J. Law. Econ.*, 1960, *3*, 1–44.

**L. Fouraker and S. Siegel**, *Bargaining Behavior*, New York 1963.

**H. Leibenstein**, "Organizational or Frictional Equilibria, X-Efficiency, and the Rate of Innovation," *Quart. J. Econ.*, Nov. 1969, *83*, 600–23.

**C. Plott and V. Smith**, "An Experimental Examination of Two Exchange Institutions," California Inst. of Tech. 1975.

**H. Shapley**, *Of Stars and Men*, Boston 1964.

**R. Sherman**, "The Psychological Difference Between Ambiguity and Risk," *Quart. J. Econ.*, Feb. 1974, *88*, 166–69.

**M. Shubik**, *Strategy and Market Structure*, New York 1959.

———, "A Trading Model to Avoid Tatonnement Metaphysics," Cowles Foundation disc. pap. no. 368, Feb. 13, 1974.

**S. Siegel**, "Decision Making and Learning under Varying Conditions of Reinforcement," *Ann. N. Y. Acad. Sci.*, 1961, *89*, 766–83.

——— **and L. Fouraker**, *Bargaining and Group Decision Making*, New York 1960.

**V. L. Smith**, "An Experimental Study of Competitive Market Behavior," *J. Polit. Econ.*, Apr. 1962, *70*, 111–37.

———, "Effect of Market Organization on Competitive Equilibrium," *Quart. J. Econ.*, May 1964, *78*, 181–201.

———, "Experimental Auction Markets and the Walrasian Hypothesis," *J. Polit. Econ.*, Aug. 1965, *73*, 387–93.

———, "Notes on Some Literature in Experimental Economics," Social Science working pap. no. 21, California Inst. of Tech., Feb. 1973, 1–27.

# THE APPLICATION TO ECONOMICS OF DIFFERENTIAL TOPOLOGY AND GLOBAL ANALYSIS

# Regular Differentiable Economies

*By* GERARD DEBREU*

The recent introduction of differential topology into economics was brought about by the study of several basic questions that arise in any mathematical theory of a social system centered on a concept of equilibrium. The purpose of this paper is to present a detailed discussion of two of those questions, and then to make a rapid survey of some related developments of the last five years.

Let $e$ be a complete mathematical description of the economy to be studied (e.g., for an exchange economy, $e$ might be a list of the demand functions and of the initial endowments of the consumers). Assumptions made a priori about $e$ (e.g., assumptions of continuity on the demand functions) define the space $\mathcal{E}$ of economies to which the study is restricted. By a state of an economy we mean a list of specific values of all the relevant endogenous variables (e.g., prices and quantities of all the commodities consumed by the various consumers). We denote by $S$ the set of conceivable states. Now a given equilibrium theory associates with each economy $e$ in $\mathcal{E}$, the set $E(e)$ of equilibrium states of $e$, a subset of $S$.

As a first test of the adequacy of this mathematical model, it must be possible to prove that for every element $e$ of a sufficiently broad class $\mathcal{E}$, the set $E(e)$ is not empty. This is the existence problem that has been extensively studied during the last decades. Mentioning only the early contributions of John von Neumann and Abraham Wald, I refer to the comprehensive survey of the literature by Kenneth Arrow and F. H. Hahn. The mathematical tools for the solution were provided by

* Department of Economics, University of California, Berkeley. I gratefully acknowledge the support of the National Science Foundation in the preparation of this article.

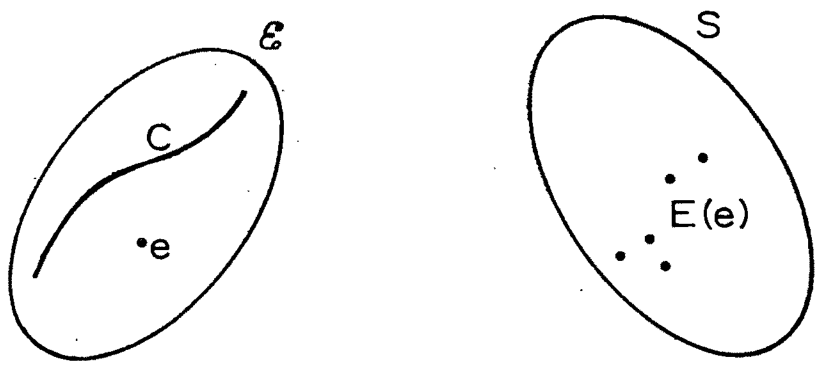



FIGURE 1

algebraic topology in the form of fixed point theorems, or directly related results. An alternative approach of the last ten years, associated mainly with the name of Herbert Scarf, has consisted of developing efficient algorithms for the computation of an approximate equilibrium. The mathematical techniques used here were those of combinatorial topology.

Having obtained a general solution to the existence problem, one must investigate the structure of the set $E(e)$ of equilibria of the economy $e$. Consider an economy that has an equilibrium $x$ such that in any neighborhood of $x$, there are infinitely many other equilibria. In this situation the explanation of equilibrium is essentially indeterminate, at least near $x$. Moreover the economic system $e$ is unstable in the sense that arbitrarily small perturbations from $x$ to a neighboring equilibrium induce no tendency for the state of the economy to return to $x$. It is therefore highly desirable to have an economy $e$ for which the set of equilibria is *discrete*, i.e., such that for every equilibrium $x$ in $E(e)$, there is a neighborhood of $x$ in which $x$ is the unique equilibrium of $e$. Unfortunately, even if every agent in the economy $e$ is mathematically very well-behaved, one may obtain a set $E(e)$ that is not discrete (in the simplest case of an exchange economy with two consumers and two commodities, one can easily find in the associated Edgeworth box a set $E(e)$ made up of a continuum of points). The pathology is due to the manner in which the agents are matched, a situation entirely different from that of existence theory where it was possible to give general conditions on the behavior of each agent separately ensuring that the set $E(e)$ would not be empty.

The way out of this difficulty is provided by differential topology. It consists of making suitable differentiability assumptions on the functions entering the description of $e$ (e.g., demand functions are assumed to be continuously differentiable), and to define a concept of *regular* economy such that (a) the critical set $C$ of nonregular economies is a negligible subset of $\mathcal{E}$, and (b) every regular economy has a discrete set of equilibria.

Actually an adequate model $e$ of the economy must have still another property. Specifically if $e'$ is close to $e$, then one would like the set of equilibria $E(e')$ to be close to $E(e)$. Otherwise an arbitrarily small error in the determination of the characteristics of $e$ would yield an entirely different set of equilibria, thus depriving the theory of much of its explanatory power. Therefore it is also desirable for the definition of regularity to be such that (c) in a neighborhood of a regular economy, the set of equilibria depends continuously on the economy. The questions of discreteness of the set of equilibria and of continuous dependence of the set of equilibria on the economy, or closely related questions, have a long history in the study of physical systems. The recent work of R. Thom has considerably extended the range of their applications, in particular to biological systems. These questions clearly have no less relevance for social systems.

The solution of the problem that has just been outlined rests on A. Sard's theorem which Thom once characterized as one of the three main results of Mathematical Analysis. Consider a continuously differentiable function $f$ from the real line $R$ to $R$, and define a *critical point* as a point $x$ where the derivative of $f$ vanishes. The set of critical points of $f$ can obviously be large. In the extreme case of a constant function it is the whole of $R$. Define now a *critical value* of $f$ as the image of a critical point. In Figure 2 three critical values $y^1$, $y^2$, $y^3$ are displayed.

One feels that the set of those critical values is necessarily small, and indeed Sard's theorem asserts that it is negligible. To be precise, it has (Lebesgue) *measure*

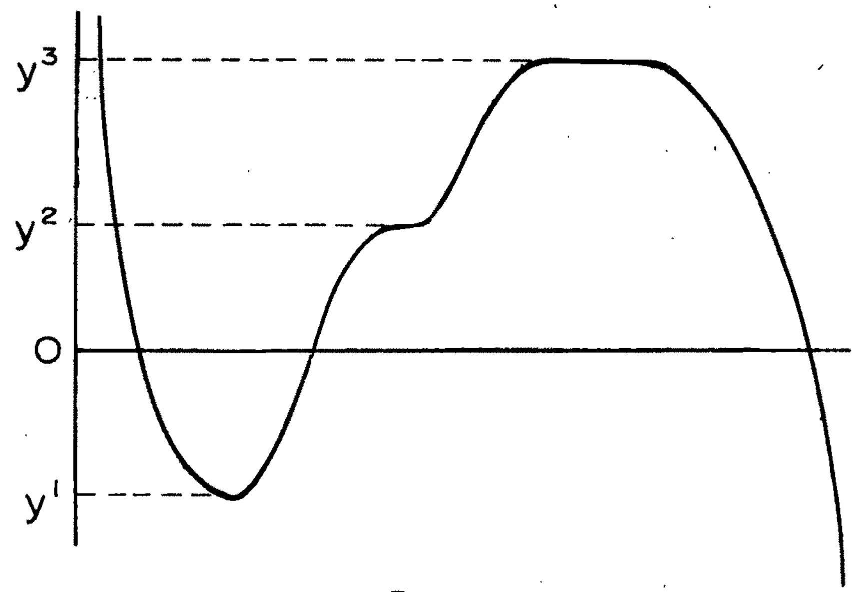



FIGURE 2

*zero*. In other words, given an arbitrarily small positive number $\epsilon$, one can find a countable collection of intervals such that their union covers the critical set, and that the sum of the lengths of these intervals is smaller than $\epsilon$. Sard's theorem holds as well for a continuously differentiable function $f$ from an $m$-dimensional Euclidean space to an $m$-dimensional Euclidean space. Here a *critical point* of $f$ is a point where the determinant of the Jacobian of $f$ vanishes. As before, a *critical value* of $f$ is the image of a critical point, and the set of critical values has measure zero, i.e., it can be covered by a countable collection of $m$-dimensional cubes of arbitrarily small total volume. A *regular value* of $f$ is, by definition, a noncritical value.

Sard's theorem is valid in still more general conditions. In order to present a stronger version that we will use later on, we need the concept of a (differentiable) manifold. A function $g$ from an $n$-dimensional Euclidean space to an $n$-dimensional Euclidean space is said to be a *differentiable isomorphism* if it is one-to-one, and if $g$, as well as its inverse $g^{-1}$ are continuously differentiable. Differential topology studies properties that are invariant under differentiable isomorphisms, a principle

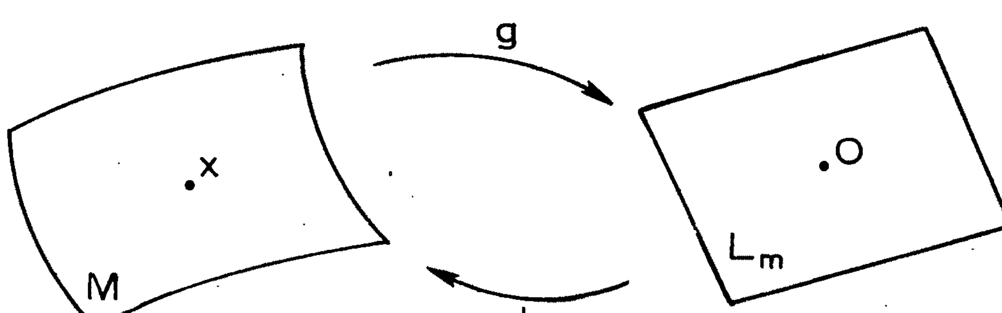



FIGURE 3

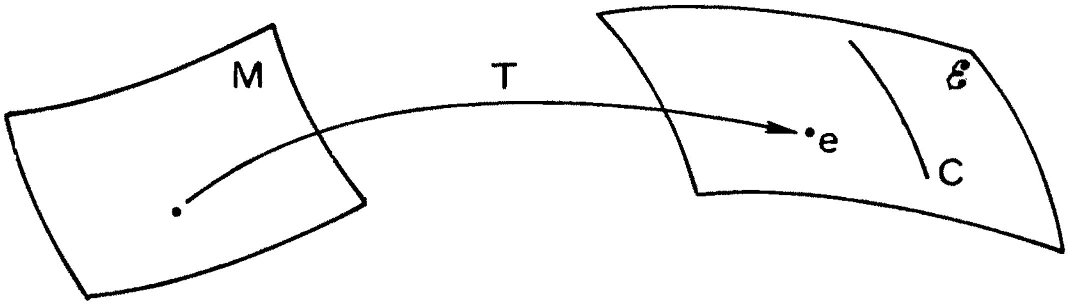



FIGURE 4

that leads to the following definition. A subset $M$ of a Euclidean space $L$ is an $m$-dimensional (differentiable) *manifold* if for each point $x$ of $M$ there is an $m$-dimensional linear subspace $L_m$ of $L$ such that a neighborhood in $L$ of $x$ and the part of $M$ that it contains are differentiably isomorphic to a neighborhood in $L$ of $O$ and the part of $L_m$ that it contains. In other words, at every one of its points, $M$ is locally an $m$-dimensional Euclidean space up to a differentiable isomorphism. Clearly one can do differential calculus on manifolds exactly as on Euclidean spaces; one can also easily define a subset of measure zero of a manifold $M$. Then the statement of Sard's theorem becomes: if $f$ is a continuously differentiable function from an $m$-dimensional manifold $M_1$ to an $m$-dimensional manifold $M_2$, the set of critical values of $f$ has measure zero in $M_2$.

To illustrate these general principles, assume that the economy $e$ can be characterized by finitely many real parameters, or more precisely, that $\mathcal{E}$ is a finite-dimensional manifold. A simple strategy for showing that almost every economy in $\mathcal{E}$ is well-behaved consists of introducing a manifold $M$ of the same dimension as $\mathcal{E}$, and a continuously differentiable function $T$ from $M$ to $\mathcal{E}$ such that a *regular* economy, defined as a regular value of $T$, actually has properties (b) and (c). Sard's theorem implies that the set $C$ of critical economies has measure zero.

To give an even more specific illustration, we consider an exchange economy with $l$ commodities, $m$ consumers, and fixed demand functions. Therefore the parameters of the economy are the initial holdings of each commodity by each consumer, a list $e$ of $lm$ positive numbers. We denote by $\mathcal{E}$ the set of those $lm$ lists, and we observe for later use that the dimension of $\mathcal{E}$ is $lm$. The state of the economy is taken to be the price-system, i.e., a list $p$ of $l$ positive numbers. Since multiplying all prices by the same positive number does not affect the behavior of the agents, we can normalize the price-system and restrict it to belong to a manifold $S$, such that dim $S=l-1$, for instance $S$ may be the positive part of the unit sphere in the $l$-dimensional Euclidean space. Given the economy $e$ and the price-system $p$, we write the $l$-list of the excesses of demand over supply on every market as $F(e, p)$, a vector in the $l$-dimensional space $R^l$. The price-system $p$ is an equilibrium state if and only if

$$F(e, p) = 0. \tag{1}$$

Given $e$, the set of $p$ satisfying (1) is $E(e)$. However, the excess demand function $F$ obeys Walras' law. Namely for every $e$ and $p$, the value of the excess demand equals 0; i.e., $p \cdot F(e, p)=0$. Consequently the equilibrium condition (1) is equivalent to equating to $O$ the list $\hat{F}$ of the first $l-1$ components of $F$,

$$\hat{F}(e, p) = 0. \tag{2}$$

Now if the individual demand functions are continuously differentiable, we can follow the strategy we have outlined above in several ways. An elementary treatment that does not use the concept of a manifold can be given as in Debreu (1970, 387–392). An alternative, and more satisfactory, solution takes as a central concept the set $M$ of pairs $(e, p)$ in the Cartesian product $\mathcal{E} \times S$ satisfying (2) (Stephen Smale 1974, pp. 1–14, Y. Balasko 1975, pp. 95–118). The space $\mathcal{E} \times S$ is of dimension $lm+l-1$, and the equilibrium condition (2) imposes $(l-1)$ restrictions on the pair $(e, p)$. Therefore one may expect $M$ to be a manifold of dimension $lm$, i.e., of the same dimension as $\mathcal{E}$. Indeed this is easily shown to be the case. The function $T$ from the *equilibrium manifold* $M$ to $\mathcal{E}$ could hardly be simpler; it is the transformation $(e, p) \mapsto e$, i.e., the projection from $M$ into $\mathcal{E}$. A regular economy is then defined as a regular value of $T$. Equivalently $e$ is a regular economy if for every $(e, p)$ in $M$ projecting into $e$, the projection of the tangent space of $M$ at $(e, p)$ covers $\mathcal{E}$. Three critical economies $e^1$, $e^2$, $e^3$ are displayed on Figure 5. Sard's theorem yields the conclusion that (a) the set $C$ of critical economies has measure zero.

Given a regular economy $e$ in $\mathcal{E}$ (as on Figure 5), one obtains the set $E(e)$ of equilibrium price-systems $p$ by taking in $M$ the set $T^{-1}(e)$ of inverse images of $e$ by the projection $T$, and by projecting $T^{-1}(e)$ into $S$. At every point of $T^{-1}(e)$, the determinant of the Jacobian of $T$ is different from $O$, and a direct application of the inverse function theorem yields that $T^{-1}(e)$ is a discrete subset of $M$. The pathologies associated with critical economies are made clear by Figure 5. For instance the economy $e^1$ has a discrete set of (two) equilibria, but a continuous displacement of the economy in a neighborhood of $e^1$ produces at $e^1$ a sudden change in the set of equilibria. In particular, consider $e$ on the right of $e^1$ and the associated equilibrium $p$ on the upper branch of $M$ (as on Figure 5). When $e$ moves left, follow $p$ by continuity. As $e$ crosses $e^1$, the equilibrium state of the system jumps to the lower branch of $M$. The economy $e^3$ has a continuum of equilibria and an isolated equilibrium. At the point $e^3$ one also observes a sudden change in the set of equilibria for a continuous displacement of the economy in a neighborhood of $e^3$. Although the economy $e^2$ is critical, it

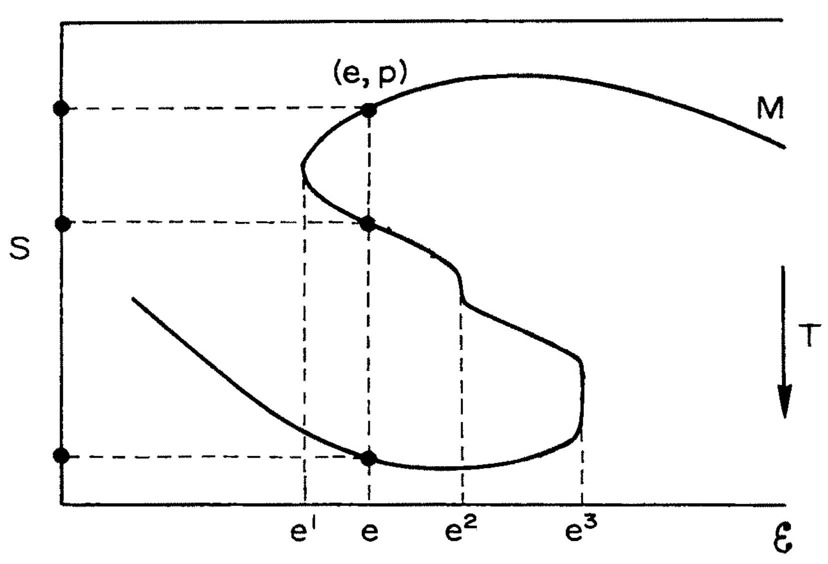



FIGURE 5

has a discrete set of (three) equilibria, and in a neighborhood of $e^3$, the set of equilibria depends continuously on the economy.

One can obtain considerably stronger conclusions by making suitable assumptions on the behavior of the excess demand function $F$ near the boundary of $S$. For instance, assume that given $e$, the excess demand $F(e, p)$ is unbounded when the price of a commodity tends to zero. In this case the set $C$ of critical economies is *closed* in addition to being of measure zero, and is therefore negligible in a strong sense. Moreover ($b'$) every regular economy has a *finite* set of equilibria, and ($c'$) in a neighborhood of a regular economy, the set of equilibria depends in a *continuously differentiable* manner on the economy.

Property ($c'$), which strengthens property ($c$) of continuous dependence of the set of equilibria $E(e)$ on the economy $e$, has made it possible to answer the question of the rate of convergence of the core of an economy. Many authors from Edgeworth to W. Hildenbrand have shown with increasing generality that the core and the set of Walras equilibria of an economy tend toward each other as the number of agents increases in such a way that every one of them becomes insignificant. This precise formulation of the idea that an economy tends to become more competitive under these conditions raises the question of the rate of that convergence. A conjecture of Shapley (1975, pp. 345–51) has led to the proof (Debreu, 1975, pp. 1–7; B. Grodal, pp. 171–186) that outside the negligible critical set, the core and the set of Walras equilibria converge to each other at least as fast as the reciprocal of the number of agents converges to zero.

The results on negligible sets of critical economies that we have presented have been considerably extended by Balasko, Chichilnisky, Delbaen, E. and H. Dierker, Fuchs, K. Hildenbrand, Ichiishi, Kalman, Laroque, Mas-Colell, Mitiagin, Rader, Schecter, Smale, and Varian. In particular, the demand functions of the consumers, which we assumed to be fixed, have been treated as variables; economies with many agents, or with many commodities have been taken into account; production has been introduced into the model; demand functions have been replaced as primitive concepts by utility functions or by preference relations, neither of which are restricted to satisfy convexity assumptions. However in most of these extensions the description of the economy requires infinitely many real parameters, i.e., the dimension of the space $\mathcal{E}$ of economies is infinite, and it is no longer possible to use the previous measure theoretical definition of a negligible set. One can now prove only that in the space $\mathcal{E}$, the critical economies form a *closed subset $C$ whose interior is empty.*

The introduction of differential topology into economic theory has made necessary a reexamination of several classical problems, for instance that of differentiable preference relations. In particular, Smale has played a leading role in extending, simplifying, and making more rigorous the study of the set of Pareto optima, and of dynamic processes converging to that set, a program of research in which de Melo, Ong, Simon, Titus, and Y. H. Wan have participated. But of special importance among the applications of Global Analysis to economics has been Smale's recent work on Global Newton Methods where he gives a differential analog of the Scarf-Eaves algorithm for the computation of an economic equilibrium, unifying their approach with the traditional dynamic economic processes of Samuelson, and of Arrow, Block, and Hurwicz.

REFERENCES

*Note:* These references include a bibliography of published work on the applications of differential topology to economics that is intended to be complete.

K. J. Arrow and L. Hurwicz, "On the Stability of the Competitive Equilibrium, I," *Econometrica*, 1958, *26*, 522–52.

———, **H. D. Block and L. Hurwicz,** "On the Stability of the Competitive Equilibrium, II," *Econometrica,* 1959, *27,* 82–109.

——— **and F. H. Hahn,** *General Competitive Analysis,* San Francisco 1971.

**Y. Balasko,** "Some Results on Uniqueness and on Stability of Equilibrium in General Equilibrium Theory," *J. Math. Econ.,* 1975, *2,* 95–118.

———, "The Graph of the Walras Correspondence," *Econometrica,* 1975, *43,* 907–12.

**G. Debreu,** "Economies with a Finite Set of Equilibria," *Econometrica,* 1970, *38,* 387–92.

———, "Smooth Preferences," *Econometrica,* 1972, *40,* 603–15.

———, "Four Aspects of the Mathematical Theory of Economic Equilibrium," *Proceedings of the International Congress of Mathematicians, Vancouver,* 1974, 65–77.

———, "The Rate of Convergence of the Core of an Economy," *J. Math. Econ.,* 1975, *2,* 1–7.

**E. Dierker,** "Two Remarks on the Number of Equilibria of an Economy," *Econometrica,* 1972, *40,* 951–53.

———, *Topological Methods in Walrasian Economics,* Lecture Notes in Economics and Mathematical Systems, *92,* Berlin 1974.

——— **and H. Dierker,** "The Local Uniqueness of Equilibria," *Econometrica,* 1972, *40,* 867–81.

**H. Dierker,** "Smooth Preferences and the Regularity of Equilibria," *J. Math. Econ.,* 1975, *2,* 43–62.

———, "Equilibria and Core of Large Economies," *J. Math. Econ.,* 1975, *2,* 155–69.

**I. Ekeland,** "Topologie Différentielle et Théorie des Jeux," *Topology,* 1974, *13,* 375–88.

**G. Fuchs,** "Private Ownership Economies with a Finite Number of Equilibria," *J. Math. Econ.,* 1974, *1,* 141–58.

———, "Structural Stability for Dynamical Economic Systems," *J. Math. Econ.,* 1975, *2,* 139–54.

——— **and G. Laroque,** "Continuity of Equilibria for Economies with Vanishing External Effects," *J. Econ. Theory,* 1974, *9,* 1–22.

**B. Grodal,** "The Rate of Convergence of the Core for a Purely Competitive Sequence of Economies," *J. Math. Econ.,* 1975, *2,* 171–86.

**J. Grote,** "A Global Theory of Games 1," *J. Math. Econ.,* 1974, *1,* 223–35.

**J. Harsanyi,** "Games with Randomly Disturbed Payoffs: A New Rationale for Mixed-Strategy Equilibrium Points," *Int. J. Game Theory,* 1973, *2,* 1–23.

———, "Oddness of the Number of Equilibrium Points: A New Proof," *Int. J. Game Theory,* 1973, *2,* 235–50.

**K. Hildenbrand,** "Continuity of the Equilibrium-Set Correspondence," *J. Econ. Theory,* 1972, *5,* 152–62.

———, "Finiteness of $\Pi$ ($\mathcal{E}$) and Continuity of $\Pi$," Appendix to Ch. 2 in W. Hildenbrand, *Core and Equilibria of a Large Economy,* Princeton 1974.

**A. Mas-Colell,** "A Note on a Theorem of F. Browder," *Math. Programming,* 1974, *6,* 229–33.

———, "On the Continuity of Equilibrium Prices in Constant-Returns Production Economies," *J. Math. Econ.,* 1975, *2,* 21–33.

**W. de Melo,** "Optimization of Several Functions," in *Dynamical Systems—Warwick 1974,* Lecture Notes in Mathematics, 468, A. Dold and B. Eckmann, eds., Berlin 1975.

**B. Mitiagin,** "Notes on Mathematical Economics," *Uspehi Matematičeskih Nauk,* 1972, *27,* 3–19, translated in *Russian Mathematical Surveys.*

**J. von Neumann,** "Zur Theorie der Gesellschaftsspiele," *Mathematische Annalen,* 1928, *100,* 295–320, translated in *Contributions to the Theory of Games IV,* A. W. Tucker and R. D. Luce, eds., Princeton 1959.

———, "Über ein ökonomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunktsatzes," *Ergebnisse eines mathematischen Kolloquiums,* 1937, No. 8, 73–83, translated in *Rev. Econ. Studies,* 1945, *13,* 1–9.

**J. T. Rader,** *Theory of General Economic Equilibrium,* New York 1972, 182–201.

———, "Absolutely Continuous Constrained Maximizers," *J. Optimization Theory and Applications,* 1973, *12,* 107–28.

———, "Nice Demand Functions," *Econometrica,* 1973, *41,* 913–35.

**R. Saigal and C. Simon,** "Generic Properties

of the Complementarity Problem," *Math. Programming*, 1973, *4*, 324–35.

P. A. **Samuelson**, "The Stability of Equilibrium: Comparative Statics and Dynamics," *Econometrica*, 1941, *9*, 97–120.

———, *Foundations of Economic Analysis*, Cambridge, Mass. 1947.

A. **Sard**, "The Measure of the Critical Points of Differentiable Maps," *Bull. of the Amer. Math. Soc.*, 1942, *48*, 883–90.

H. **Scarf**, *The Computation of Economic Equilibria*, New Haven 1973.

L. S. **Shapley**, "An Example of a Slow-Converging Core," *Int. Econ. Rev.*, 1975, *16*, 345–51.

C. P. **Simon** and C. **Titus**, "Characterization of Optima in Smooth Pareto Economic Systems," *J. Math. Econ.*, 1975, *2*, 297–330.

S. **Smale**, "Global Analysis and Economics I," in *Dynamical Systems*, M. Peixoto, ed., New York 1973.

———, "Global Analysis and Economics IIA," *J. Math. Econ.*, 1974, *1*, 1–14.

———, "Global Analysis and Economics III," *J. Math. Econ.*, 1974, *1*, 107–17.

———, "Global Analysis and Economics IV," *J. Math. Econ.*, 1974, *1*, 119–27.

———, "Global Analysis and Economics V," *J. Math. Econ.*, 1974, *1*, 213–21.

———, "Optimizing Several Functions," in *Manifolds-Tokyo, 1973*, Akio Hattori, ed., Tokyo 1975.

———, "Sufficient Conditions for an Optimum," in *Dynamical Systems-Warwick 1974*, Lecture Notes in Mathematics, *468*, A. Dold and B. Eckmann, eds., Berlin 1975.

———, "An Approach to the Analysis of Dynamic Processes in Economic Systems," in *Equilibrium and Disequilibrium in Economic Theory*, G. Schwödiauer, ed., 1975.

D. **Sondermann**, "Smoothing Demand by Aggregation," *J. Math. Econ.*, 1975, *2*, 201–23.

R. **Thom**, *Stabilité Structurelle et Morphogénèse*, Reading 1972, translated as *Structural Stability and Morphogenesis*, Reading 1975.

H. R. **Varian**, "On Persistent Disequilibrium," *J. Econ. Theory*, 1975, *10*, 218–28.

———, "A Third Remark on the Number of Equilibria of an Economy," *Econometrica*, 1975, *43*, 985–86.

A. **Wald**, "Über einige Gleichungssysteme der mathematischen Ökonomie," *Zeitschrift für National Ökonomie*, 1936, *7*, 637–70, translated in *Econometrica*, 1951, *19*, 368–403.

Y. H. **Wan**, "On Local Pareto Optima," *J. Math. Econ.*, 1975, *2*, 35–42.

E. C. **Zeeman**, "On the Unstable Behavior of Stock Exchanges," *J. Math. Econ.*, 1974, *1*, 39–49.

# Dynamics in General Equilibrium Theory

By STEPHEN SMALE*

I find myself something of an outsider in writing a paper like this; and for a mathematician to write a paper with no technical mathematics is a real (but very constructive) confrontation with the problem of communication. What I hope to achieve here is to convey some of my ideas on the relation between dynamics and the traditionally static, economic equilibrium theory.

Since I have been working in mathematical economics, I have been struck by the number of attacks on general equilibrium theory, on mathematical economics and even economic theory in general. Coming from a radical background, I have much sympathy for some of the arguments brought forth.

I first want to say a few words on the subject of mathematics in economics and economic theory in general. The role of theory per se hardly requires defense; theory can give a deeper understanding of any subject, subtle relations are seen, inconsistent ideas are exposed, new horizons are revealed.

A criticism commonly made of economic theory is its failure to make predictions of crises in the country or to anticipate correctly unemployment or inflation. One must be cautious in the social sciences about looking toward physics for answers. However, some comparisons with the physical sciences seem profitable in connection with the above criticism. In those sciences, where theory itself is in a far more advanced state, limitations can be seen in a similar way. For example a given individual human body functions according to physical principles; however no physical scientist would predict a heart attack. The physical theory gives understanding of aspects of what goes on in the human body only under very idealized conditions. The physical theories eventually play some role in the education of medical doctors, who can then say some things, some times about a patient's susceptibility to a heart attack, preventive measures, and cures.

The economy of the world or even a nation is a very complex phenomenon, like a human body, involving a number of factors, both economic and political. It is no more reasonable to expect economic theorists to predict a nation's economic future than for a theoretical scientist to predict the future health of an individual.

Questions about the need for mathematics in economic theory have been raised. Indeed, the successes of mathematics in economics have not been nearly as impressive as in physics. Yet the notion of money and prices already introduces mathematics into economics; and the mathematics becomes deeper with the equation of equilibrium, supply equals demand. When one considers the equation, supply equals demand for several interdependent markets, the mathematical problem already takes on some sophistication.

What is special about "general equilibrium theory" as opposed to economic theory in general? For me the importance

* Dept. of Mathematics, University of California, Berkeley.

of general equilibrium theory lies in its traditions which are deeper than any other part of economic theory. These traditions, which of course derive from actual economic history, explain why equilibrium theory has played such a central role and possesses such depth in content. Many of the procedures and mathematical methods used in other parts of economics grew out of developments in general equilibrium theory. Also equilibrium theory plays an important role in communication within the economics profession. Since most economists are knowledgeable in equilibrium theory, they can understand new ideas more readily when presented in that context. Also, since equilibrium theory has been studied so much, new ideas introduced there show weaknesses and also strengths most quickly.

After all of this is said, equilibrium theory will eventually stand or fall, depending on its truth as an important idealization of actual economic systems or as a model with values of justice, of efficient distribution and of utilization of resources. As a normative theory, I find great merit in its decentralization features (Schumacher's popular book, "Small is Beautiful" expresses some of my sentiment on decentralization). There are also, without doubt, basic failings in the theory which are well-presented in the economic literature, and there are some weaknesses which I wish to discuss presently. In fact these problems can be seen and understood especially clearly due to the well-developed structure of the theory; and one can use the body of general equilibrium theory as a tool in developing alternate models. To me it would seem overly difficult to construct and communicate successfully any alternate economic theory without having first studied very thoroughly equilibrium theory.

I would like to give some reasons why I feel equilibrium theory is far from satisfactory. For one thing the theory has not successfully confronted the question, "How is equilibrium reached?" Dynamic considerations would seem necessary to resolve this problem. Another weakness is the reliance of the theory on long range optimization.

In the main model of equilibrium theory, say as presented in Gerard Debreu's *Theory of Value,* economic agents make one life-long decision, optimizing some value. With future dating of commodities, time has almost an artificial role. The model is reminiscent of John von Neumann's game theory. I like to make an analogy between "Theory of Value" and the game theoretic approach to chess. The possible strategies are laid out to each player in advance, paths in a game tree, or a set of moves, one move to each position that could possibly occur. Each player makes a single choice of strategy. The strategies are compared and the game is over. Of course, chess isn't played like this. And in a situation more complicated than chess, where life-long consumption plans replace strategies, I don't believe economic agents act that way either.

In fact even the very best chess players don't analyze very many moves and certainly don't make future commitments. Their experience together with the environment at the moment (the position), some rules of thumb and some other considerations lead to decisions on the playing board.

My personal economic decisions are of a similar nature, from buying a book to buying a house; from a decision to travel to decisions about my job. Between one economic decision and another, there has been a real passage of time, circumstances have changed, and the new decision takes place in this new environment.

Long-run optimization would be im-

practical, even if it were emotionally acceptable, because of barriers of complexity. Complexity keeps us from analyzing very far ahead. The amount of time involved in making a decision is an important factor, for a chess player or a purchaser. Dynamic models based on some kind of behavioral strategy could meet these objections.

Sometimes static theories pose paradoxes whose resolution lies in a dynamic perspective. Let me give an example from the theory of duopoly of the Cournot (or Nash) equilibrium solution. Under classical hypotheses on the profit functions, consider such an equilibrium $(r_1, r_2)$ where $r_i$ is the rate of production of duopolist number $i$. Then agent 1's rate of production maximizes profit among all such rates with $r_2$ fixed. On the other hand the solution is unsatisfactory because there is another state, say, $(r'_1, r'_2)$, where each duopolist with reduced production is taking a higher profit. In the actual dynamic world it is unlikely that the duopolists would stay at the Cournot solution, knowing that they would both be better off at nearby states.

It doesn't require explicit cooperation for these agents to move off the Cournot solution. In fact with flow of information and implicit threats in the context of a passage of time, one can argue that the duopolists will move to an optimal state from a Cournot state. But this resolution requires a real passage of time, that after each market move, the opposing duopolist has another opportunity to move. One can readily think of examples of duopoly where an increase of advertising is withheld by one agent knowing the other agent would match an increase and both would be worse off. James Friedman has written on this topic.

I feel that dynamics could also play a role in the resolution of Kenneth Arrow's paradox in the politics of social choice. Politics and elections in particular are actually part of a dynamic process, balloting being just a stage. The process looked at as a whole involves a number of moves such as a candidate's speech, a political ad, revising a position on some issue, etc. After each action of a candidate, other candidates have the option of taking an action of their own; voters' opinions evolve. For this reason, I would think that a dynamical model of the political process would give much better perspective than a static model of simple voting. In relation to this, the work of G. Kramer comes to mind.

We return to the subject of equilibrium theory. The existence theory of the static approach is deeply rooted to the use of the mathematics of fixed point theory. Thus one step in the liberation from the static point of view would be to use a mathematics of a different kind. Furthermore, proofs of fixed point theorems traditionally use difficult ideas of algebraic topology, and this has obscured the economic phenomena underlying the existence of equilibria. Also the economic equilibrium problem presents itself most directly and with the most tradition not as a fixed point problem, but as an *equation*, supply equals demand. Mathematical economists have translated the problem of solving this equation into a fixed point problem.

I think it is fair to say that for the main existence problems in the theory of economic equilibria, one can now bypass the fixed point approach and attack the equations directly to give existence of solutions, with a simpler kind of mathematics and even mathematics with dynamic and algorithmic overtones. In the last part of the paper we elaborate on this point.

Behind my own work on the questions of dynamics in economics, lies certain foundational work in the equilibrium theory in terms of calculus. The early

development of mathematical economics, including the 19th century and even up to World War II, was largely in terms of calculus; it was no doubt the influence of game theory, and associated fixed point theorems that gradually reduced the dependence on calculus. In *Theory of Value* in 1959, Debreu wrote of the work of von Neumann and Oskar Morgenstern which freed mathematical economics from its traditions of differential calculus and compromises with logic. He wrote of the radical change of mathematical tools from calculus to convexity and topology.

But then in his paper on a finite number of equilibria, Debreu returned to calculus tools; my own work, "Global Analysis and Economics," has been to try to systematize the use of calculus in equilibrium theory. This can be justified on several grounds. First, the theory is brought closer to the practice. With calculus, one has in the derivative a linear approximation. It is these linear conditions that are so basic to practical economic studies. Comparative statics depend on derivatives; the same is usually true for stability conditions; dynamic questions are more accessible via calculus. When general equilibrium theory is developed on calculus mathematics, not only is theory brought closer to practice, but greater unity is achieved. Furthermore, recent work on approximation by differentiable functions in economics gives further justification to the use of calculus.

Finally before moving to the constructive side of the question of dynamics in equilibrium theory, it is worth making a remark on the nature of goods. It seems from our experience that it is important especially in modeling dynamics to put goods into two ideal classes, either completely perishable or completely durable. The theory seems different for the two kinds. For example, Walras equilibrium seems suited to the perishable, continually endowed class of goods, while for durable goods, the kind of equilibrium found in the fundamental theorem of welfare economics seems more appropriate. We hope that the rest of the paper makes this point clearer.

We discuss now the results of our paper, "Exchange Processes with Price Adjustment." This is a model of a market of durable goods; a particular example from personal experience of such a market is a weekend "mineral bourse" where agents with minerals and/or money meet to exchange, buy, and sell fine mineral specimens for collectors. Here one sees a truly dynamic process of exchange and price adjustments which converges to an equilibrium toward the end of the weekend of the "bourse."

An early work in some of the same spirit was done in 1962 by Frank H. Hahn, Hirofumi Uzawa, and Takashi Negishi. This approach is called a "nontatonment," because in place of the Walrasian Tatonment, a sequence of actual trades takes place over time. On the other hand, our model differs from the 1962 work in that it is nondeterministic, the dynamics proceeds without an ordinary differential equation, long-run optimization is dropped, and a large body of examples is constructed.

The main result of our paper can be stated as follows:

"In a pure exchange economy, an exchange price adjustment process, responsive to transaction costs, and which doesn't stop unless forced to by market conditions, converges to a price equilibrium. There exists such processes starting from any state of any (pure exchange) economy."

In the paper, mathematical content is given to all of the phrases used here, and the result is proved. Here we give a brief explanation of some of the terms used.

A "state" of an economy means a set of

data characterizing the economy at a given time. Our use of the word "state" is akin to its use in physics. For example in a pure exchange economy, a state will consist of an allocation of the resources, or equivalently the set of goods of each agent and a price system.

A state in general will change over time, e.g., by exchange and price adjustments. Thus a "process" as used in the main result, means a passage in time of a state. Or equivalently, a process is a path (over time) in the space of states of an economy.

This process, to qualify as an *exchange* process, must satisfy economically justifiable conditions for exchange which are embodied in the following axiom. The exchange axiom for the process asserts that: (a) the total resources of the economy are constant (there is no production); (b) exchange takes place at current prices; (c) an exchange increases satisfaction of the participating agents; and (d) some exchange will take place provided that it is possible consistent with (a), (b), and (c).

For the process to qualify as a price adjustment process (as in the quoted main result), we demand that it satisfy a price adjustment axiom defined in terms of a short run version of demand. A usual excess demand approach requires long-run optimization for the agents while our spirit is closer to that of behavioral strategies. At given prices and goods possessed one defines the infinitesimal demand of an agent to be the direction his preferences take him when restricted to his budget set.

The price adjustment axiom asserts that prices adjust in the direction of some weighting of the infinitesimal demands of all the agents.

A Walrasian price equilibrium depends on the traders' endowments. Thus if one allows a real passage of time, say an actual exchange to take place, and several such, this initial endowment becomes forgotten. Thus if one allows a "nontatonment" kind of time passage, one must replace a Walrasian price equilibrium by a different notion of price equilibrium.

As stated in our main result above, a price equilibrium is a feasible allocation and price system where, for each agent, satisfaction is maximized on his budget set defined relative to his wealth at equilibrium. Equivalently, a price equilibrium is an optimal allocation together with a supporting price system, as studied in the correspondence of the fundamental theorem of welfare economics.

A detailed mechanism of price setting and transactions is not developed, but it seems likely that the model is consistent with doing this explication.

Next we discuss some problems and results on the classical Walrasian model from the point of view of dynamics and algorithms. We prefer an alternate, well-known interpretation of the Walras model to that given in Debreu's *Theory of Value*.

Suppose that the goods are perishable, with labor a main example. One might envision a situation where each day an agent starts his economic activity with a fixed endowment of labor, or fish which won't keep. The next day he will have a new endowment of the same, but none left from the day before.

The consumption variables are the amounts of commodities consumed each day, of an agent. Thus both the endowment and consumption bundles in commodity space will be interpreted as the *rates* of endowment (fixed over time) and consumption respectively. A completely satisfactory dynamics (which isn't available) for this problem would construct and analyze paths over time in the space of states, that is commodity vectors for each agent and price systems (or sets of price systems). These paths should obey economically justifiable axioms of exchange and price adjustment, and

probably should lead, at least under some economic conditions to a Walrasian equilibrium, starting from an endowment allocation and any price system. At the most satisfactory level, these paths should be given interpretations in terms of individual agent's actions in price offerings and purchases. In my view, a behavioral strategy for agents would be more desirable than decisions based on long run optimization.

Martin Shubik and I have worked on the problem in this setting without any definite success. On the other hand, it seems as if we might eventually obtain such a model with convergence provided a condition such as gross substitutes is satisfied. I should emphasize that I am not talking about any "tatonment" in this kind of dynamic, but rather an actual process, where agents are adjusting their goals, consumptions, prices over time to arrive at balanced budgets where the value of the rate of consumption equals the value of the rate of the fixed endowment for each agent.

I would like to turn now to some work carried out in my article "A Convergent Process of Price Adjustment and Global Newton Methods," which has more success on the mathematical side of the above problem.

One way of looking at this work is to first alter the Scarf Algorithm from finding fixed points to solving a system of equations, especially the system, supply equals demand in many variables. We define an ordinary differential equation, called a "global Newton," which is a version of (the altered) Scarf's Algorithm. Under rather general hypotheses (comparable to those needed in the execution of Scarf's Algorithm), solutions of the global Newton converge to the set of solutions of the original system (e.g., supply equals demand). Combining this fact with methods of numerical analysis, one obtains a different but analogous algorithm to that of Scarf. Morris Hirsch and I have implemented this effectively on a computer and are developing the algorithm from a numerical analysis point of view. It applies to systems of $n$ nonlinear equations in $n$ variables, without hypotheses on the system of nonzero Jacobian, convexity, or monotonicity.

Let $z(p)$ be the excess demand as a function of prices $p=(p_1 \ldots p_l)$ so that $p^*$ is an equilibrium if $z(p^*)=0$. Then the global Newton takes the form

$$(1) \qquad Dz(p)\frac{dp}{dt} = -\lambda z(p),$$

$$\operatorname{sign} \lambda = \operatorname{sign} \text{Determinant } Dz(p)$$

Here $Dz(p)$ is the matrix of first partial derivatives of $z$. If one takes $\lambda=1$ and uses Euler's method of discrete approximation to (1) then one obtains Newton's method for solving $z(p)=0$. Using equation (1), one can obtain a proof of the existence of economic equilibrium without using fixed point theorems or algebraic topology.

Consider the problem of representing a process of price adjustments by (1). Recall that the classical "tatonment" process has an embodiment in the equation

$$(2) \qquad \frac{dp}{dt} = z(p)$$

Arrow, Leonid Hurwicz and Herbert Block have shown that solutions of (2) converge to economic equilibrium under hypotheses on $z$ such as gross substitutes. These hypotheses are substantial and are strong enough to imply the existence of a unique equilibrium. On the other hand Scarf subsequently showed that under classical properties on preference relations, almost all solutions of (2) could oscillate for all time.

Now (1) can be considered as a modification of (2), which involves more subtle intermarket relations and which will con-

verge when (2) doesn't. In particular (1) will converge in the Scarf example.

Generally speaking, (1) converges starting from almost any initial price system on the boundary of the price simplex. One could also formulate the equation in quantity space and look for an interpretation of the process in terms of budget balancing actions on the part of the agents. The interested reader could pursue these topics further in the papers cited.

## REFERENCES

K. **Arrow,** *Social Choice and Individual Values,* New Haven 1963.

——— **and L. Hurwicz,** "The Stability of the Competitive Equilibrium I," *Econometrica,* 1958, *26,* 522–52.

———, **Block and L. Hurwicz,** "The Stability of Competitive Equilibrium II," *Econometrica,* 1959, *27,* 82–109.

G. **Debreu,** *Theory of Value,* New York 1959.

———, "Economies with a Finite Set of Equilibria," *Econometrica,* 1970, *38,* 387–92.

J. **Friedman,** "A Non-Cooperative Equilibrium for Super Games," *Rev. Econ. Studies,* 1971, *113,* 1–12.

H. **Scarf,** "Some Examples of Global Instability of the Competitive Equilibrium," *Int. Econ. Rev.,* 1960, *1,* 157–72.

———, *The Computation of Economic Equilibria,* New Haven 1973.

S. **Smale,** "Global Analysis and Economics, IIA-VI," *J. Math. Econ.,* 1974–75, *1,* 1–14, 107–17, 119–27, 213–21.

———, "A Convergent Process of Price Adjustment and Global Newton Methods," preprint, Berkeley.

———, "Exchange Processes with Price Adjustment," preprint, Berkeley.

# NEW DEVELOPMENTS IN LABOR MARKET DYNAMICS

# Labor Force Participation and Earnings in a Demographic Model of the Labor Market

*By* RICHARD S. TOIKKA AND CHARLES C. HOLT*

A structural model of the labor market should improve our understanding of and ability to predict unemployment and wage change. With such a model of the labor market and the inflation process we should be better able to prescribe policies for aggregate demand and for structural improvements that are designed to reduce inflation and unemployment.

Because labor market dynamics are fast, the duration of unemployment being only a few months, a structural model to reflect turnover and search behavior must have a very short period—a quarter would be too long. However, since real wages change sluggishly, the usual allocation variables will not be fully effective in regulating the turnover and search behavior of workers and employers. Hence, we expect that the short-term dynamics of the labor market will be regulated primarily by the *availability* in the market of jobs and workers. The composition of unemployment shows wide disparities for different demographic groups which suggests the desirability of separately estimating their search and turnover behaviors.

## I. The RASST Model and Its Extensions

A group at The Urban Institute has estimated, tested and applied a structural model of the labor market, Race Age Sex Search Turnover (*RASST*), that distinguishes for each of sixteen age-race-sex groups six monthly flows between employment, unemployment and out-of-the-labor force.[1] This model is driven by exogenous jobs and the population sizes of the groups. Its transition probabilities are governed by the cyclical availability measured by the vacancy to unemployment ratio, seasonal dummies, and time trends.

This model has been used in 1974 to predict the composition of unemployment, employment, and labor participation for the recession year 1975, and more recently for 1976 based on administration forecasts of aggregate demand conditions. *RASST* also has been used to define the concept of "jobless rate" obtained by subtracting current employment from full employment labor force, which is based on current demographic composition and behavior. (See Smith and Vanski, 1975.)

Although the *RASST* model has performed well, it requires important extensions in order to be of maximum relevance for the analytic and policy uses discussed above. The extensions which we discuss here fall into the following two areas:

* The Urban Institute.

[1] For reference see Ralph E. Smith *et al.* (1974), Smith (1974), and Smith and J. E. Vanski (1975). Somewhat parallel work has been done in Canada. See Frank T. Denton *et al.*, (1975a) and (1975b).

(1) The transition probabilities that determine the search-turnover flows and in turn the stock of unemployed workers are, of course, not allocated by availability alone. Real wages and family earnings also affect worker decisions particularly to enter and leave the labor force, and very likely to quit low paid jobs. These effects should, if possible, be added to the model.

(2) To generate these wages and earnings for the demographic groups and to contribute to the explanation of wages in the inflation process, additional relations are needed. The explanatory variables should include price level changes (or better yet expectations), demographic and aggregate unemployment, discouraged worker (possibly by use of the jobless rate), relative wage levels, and proxies for family income (earned and unearned). Introducing earning as well as wage rates would require the model to generate an hours worked variable per month to complement employment, which already is generated.

## II. A Dynamic Theory of Labor Force Participation

A dynamic behavioral theory of labor force participation should begin with an analysis of household behavior. Decisions by family members determine labor force withdrawal and labor force entry. Aggregation of these discrete events at a point in time produces flows of persons changing labor force participation status.

A household may be thought to consist of a group of persons who make joint decisions to divide their time between market work. job search, and leisure or homework. Since our empirical analysis uses the monthly Current Population Survey (*CPS*) data on labor force status, it is convenient at the outset to make our theoretical constructs consistent with the *CPS* data. A person who spends any time at paid work will be considered to be "employed" ($E$). A person who spends no time at paid work but devotes some time to job search will be considered to be "unemployed" ($U$). Finally, a person who spends no time at paid work or search will be considered "not in the labor force" ($N$). Using these three labor force categories we characterize the labor market as a three-state Markov probability process.

The labor force participation dimensions of this Markov process are summarized in three transition probabilities: labor force entrances, labor force withdrawals by the unemployed, and labor force withdrawals by the employed. In general the costs and benefits of making a labor market transition are conditional on the person's current labor force status. For example, the cost of leaving the labor force is likely to be less for an unemployed worker than for an employed worker. This observation is borne out by empirical data from *CPS* which indicate that unemployed persons leave the labor force at a rate which is 7 times that for employed persons.[2]

The theory which is appropriate for analyzing the choice made by persons and households in this Markov process framework is Markov decision theory. A person's decision to change or not to change his labor force state is the outcome of a dynamic programming problem in which future utilities are dependent on the state occupied and decision variables.[3]

The structure of relative utilities will depend on relative preferences for nonmarket activities and earnings opportunities in the market sector. Preferences for

[2] This figure varies dramatically by demographic group. See Smith (1974).

[3] For a treatment of this decision problem in a dynamic programming framework, see Toikka (1975).

nonmarket activities will depend on a person's age, sex, family situation, etc. In our empirical analysis we stratify our sample by age, race, and sex to capture differences in preferences for nonmarket activity. Since opportunities for nonmarket activity are to some extent stochastic, being influenced by such factors as births, deaths, illness, and family separations, the utility of nonmarket activity has a random component which may account for frequent changes in an individual's labor force participation.

The utility of nonmarket activity $U(N)$ will be specified as a stochastic function of seasonal dummies $(D_1, \ldots, D_{11})$, and family income $(Y)$. The family income variable is positively correlated with the utility of nonmarket activity if household production is a normal good. Thus,

$$(1) \quad U(N) = b_0 + b_1 Y + \sum_i \alpha_i D_i + \epsilon \qquad b_1 > 0$$

The earnings from market work $(W)$ will be used as a measure of the utility of market activity. In the simplest kind of model, a person changes labor force participation status when the resulting change in utility net of transition costs is positive.

The probability of a person leaving the labor force when employed should vary inversely with earnings from market work and directly with the utility of nonmarket activity, so that:

$$(2) \qquad P(EN) = g[\overset{(-)}{W}, \overset{(+)}{U(N)}]$$

where (signs) denote the signs of the partial derivatives. Substituting (1) into (2) we get:

$$(3) \quad P(EN) = g[D_1, \ldots, D_{11}, \overset{(-)}{W}, \overset{(+)}{Y}]$$

Similar considerations apply with respect to the decisions to enter the labor force and to leave it when unemployed. In an expected utility maximization framework, these decisions are made by comparing the expected utility of job search with the expected utility of nonmarket activity. The probability of labor force entry $P(NL)$ varies directly with the (expected) utility of job search $U(S)$ and inversely with the (expected) utility of nonmarket activity $U(N)$. Thus,

$$(4) \qquad P(NL) = e[\overset{(+)}{U(S)}, \overset{(-)}{U(N)}]$$

Conversely, the probability of labor force withdrawal by the unemployed varies directly with the utility of nonmarket activity and inversely with the expected utility of search:

$$(5) \qquad P(UN) = d[\overset{(-)}{U(S)}, \overset{(+)}{U(N)}]$$

If we denote the probability of a job seeker getting an acceptable job in a month by $p$ and denote the utility of unemployed search by $C$, the expected utility of search can be expressed by:

$$(6) \qquad U(S) = pW + (1 - p)C$$

Substituting (1) and (6) into (4) and (5) gives us the following expressions for the labor force entry and withdrawal probabilities:

$$(7) \quad P(NL) = e\left[pW + (1 \overset{(+)}{-} p)C, \; b_0 + b_1 Y \overset{(-)}{+} \sum_i \alpha_i D_i\right]$$

$$(8) \quad P(UN) = d\left[pW + (1 \overset{(-)}{-} p)C, \; b_0 + b_1 Y \overset{(+)}{+} \sum_i \alpha_i D_i\right]$$

If the utility of unemployed search $(C)$ is constant, then the following are equivalent formulations for (7) and (8):

$$(9) \qquad P(NL) = e(D_1, \ldots, D_{11}, \overset{(+)}{pW}, \overset{(-)}{p}, \overset{(-)}{Y})$$

$$(10) \quad P(UN) = d(D_1, \ldots, D_{11}, \overset{(-)}{pW}, \overset{(+)}{p}, \overset{(+)}{Y})$$

TABLE 1—LABOR FORCE PARTICIPATION TRANSITIONS: WHITE FEMALES BY AGE

| Independent Variable | Quits to Leave the Labor Force: $P(EN)$ | | | | Labor Force Dropouts by the Unemployed: $P(UN)$ | | | | Labor Force Entrances: $P(NL)$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 16–19 | 20–24 | 25–59 | 60+ | 16–19 | 20–24 | 25–59 | 60+ | 16–19 | 20–24 | 25–59 | 60+ |
| Constant | 0.48 | 0.17 | 0.14 | 0.16 | −1.41 | −1.46 | −0.81 | −1.94 | 0.53 | 0.19 | 0.06 | 0.03 |
| | (7.6) | (5.7) | (8.4) | (3.6) | (−2.3) | (−2.5) | (−1.45) | (−1.70) | (3.1) | (1.2) | (1.1) | (1.4) |
| $W_{-1}$ | −0.10 | −0.10 | −0.13 | −0.18 | | | | | | | | |
| | (−1.3) | (−2.8) | (−6.8) | (−3.4) | | | | | | | | |
| $pW_{-1}$ | | | | | −3.08 | −2.24 | −1.85 | −5.24 | 0.40 | 0.22 | 0.05 | 0.00 |
| | | | | | (−4.3) | (−3.5) | (−2.4) | (−3.0) | (4.2) | (2.4) | (1.5) | (0.3) |
| $Y_{-1}$ | −0.12 | 0.33 | 0.55 | 0.84 | 3.71 | 3.54 | 2.13 | 4.58 | −0.89 | −0.25 | −0.04 | −0.04 |
| | (−0.30) | (1.6) | (5.1) | (2.9) | (3.0) | (3.0) | (1.8) | (2.0) | (−2.5) | (−0.8) | (−0.3) | (−0.8) |
| $p$ | | | | | 8.40 | 6.18 | 5.39 | 14.6 | −1.03 | −0.56 | −0.13 | −0.01 |
| | | | | | (4.3) | (3.6) | (2.6) | (3.0) | (−4.1) | (−2.4) | (−1.6) | (−0.3) |
| $R^2$(Corrected) | .92 | .80 | .87 | .61 | .62 | .29 | .50 | .05 | .95 | .80 | .89 | .54 |
| Durbin-Watson | 2.07 | 1.87 | 2.15 | 1.40 | 1.59 | 1.63 | 1.91 | 1.52 | 2.20 | 1.89 | 2.13 | 1.94 |

## *Empirical Results*

In Tables 1 and 2 we report some regression results for white males and females in four age groups. We estimated linear functions corresponding to (3), (9), and (10). The transition probabilities $P(EN)$, $P(UN)$, $P(NL)$ were estimated for each group by taking the *CPS* monthly gross flow of persons changing from one state to another and dividing it by the number of persons in this state of origin in the previous month.[4] The monthly real wage was estimated by the *BLS* average hourly earnings of nonsupervisory personnel divided by the Consumer Price Index.[5] As a proxy for monthly family income, we used the monthly real wage multiplied by the employment of white prime-age males and divided by white population. The series on population was linearly interpolated from annual census data. This family income proxy captures the effect of job loss by prime age males on family income. Eleven monthly dummy variables were included to remove seasonal variation.[6]

The probability of job search success ($p$) was estimated from the disaggregated gross change data. In the entry function, $p$ was estimated by the ratio of the number of persons moving directly into employment from outside the labor force to the total number entering the labor force in any month. In the withdrawal from unemployment function the estimator was the ratio of the number of persons moving into employment from unemployment to the number of unemployed in the previous month.[7]

[4] Since we do not have information on whether job separations are voluntary quits or involuntary layoffs or dismissals, the transition probability $P(EN)$ is not a pure estimate of quits to leave the labor force. However, $P(EN)$ is likely to be dominated by voluntary job separations, since persons who experience involuntary job separations are very likely to be counted as unemployed rather than out of the labor force for at least two reasons. First, the unemployment insurance system creates an incentive for persons losing jobs to classify themselves as in the labor force since benefits are conditioned on job search activity. Second, the *CPS* counts as unemployed anyone on layoff from a job if he is available for work regardless of the extent of his job search activity.

[5] This is an aggregate wage rather than a demographically specific wage. Unfortunately, there is no monthly series on wages by demographic characteristics.

[6] The seasonal dummies were constrained so that the constant term is the average of the effects over all twelve months rather than the effect in a particular month.

[7] Although the probability of job search success ($p$) is conceptually the same in the entry and withdrawal functions, we decided to measure it differently to take advantage of the additional information available in the number of persons who are recent entrants into the labor force. Presumably, the entrance decision is affected more by the probability of job search success at initial entry than by the probability of success after having been unemployed for a while. In addition, new entrants may have different characteristics and face different opportunity costs. For example, new entrants are less likely to be collecting unemployment insurance.

TABLE 2—LABOR FORCE PARTICIPATION TRANSITIONS: WHITE MALES BY AGE

| Independent Variable | Quits to Leave the Labor Force: $P(EN)$ 16–19 | 20–24 | 25–59 | 60+ | Labor Force Dropouts by the Unemployed: $P(UN)$ 16–19 | 20–24 | 25–59 | 60+ | Labor Force Entrances: $P(NL)$ 16–19 | 20–24 | 25–59 | 60+ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Constant | 0.44 | 0.12 | −0.00 | −0.00 | 0.28 | −0.29 | −0.29 | −0.98 | −0.04 | 0.45 | 0.73 | −0.01 |
| | (6.7) | (4.8) | (−1.6) | (−0.0) | (0.4) | (−0.6) | (1.7) | (−1.6) | (−0.1) | (1.2) | (3.3) | (−0.2) |
| $W_{-1}$ | −0.25 | −0.07 | −0.00 | −0.13 | | | | | | | | |
| | (−3.3) | (−2.2) | (−1.1) | (−5.0) | | | | | | | | |
| $pW_{-1}$ | | | | | −0.86 | −0.45 | −0.15 | −2.67 | 0.06 | 0.30 | 0.20 | −0.01 |
| | | | | | (−1.1) | (−1.0) | (−0.5) | (−2.7) | (0.3) | (1.5) | (1.5) | (−0.6) |
| $Y_{-1}$ | 0.76 | 0.18 | 0.04 | 0.83 | −0.10 | 0.88 | 1.01 | 2.51 | 0.41 | −0.47 | −1.23 | 0.07 |
| | (1.7) | (1.1) | (2.2) | (5.6) | (−0.7) | (0.9) | (1.8) | (2.1) | (0.6) | (−0.6) | (−2.6) | (0.7) |
| $p$ | | | | | 2.70 | 1.23 | 0.51 | 7.45 | −0.11 | −0.84 | −0.52 | 0.05 |
| | | | | | (1.3) | (1.0) | (0.6) | (2.7) | (0.2) | (−1.5) | (−1.6) | (0.8) |
| $R^2$(Corrected) | .94 | .89 | .37 | .64 | .53 | .01 | .32 | .07 | .96 | .95 | .38 | .38 |
| Durbin-Watson | 1.96 | 1.66 | 1.95 | 1.82 | 1.53 | 1.82 | 1.70 | 1.87 | 2.16 | 1.89 | 1.59 | 2.41 |

The tables contain the following data: the regression coefficients (except for those on the monthly dummies) with t-statistics in parenthesis, the $R^2$ corrected for degrees of freedom, and the Durbin-Watson statistic.

The model of labor force participation suggested above is strongly supported by the data. Of the coefficients on the economic variables, 58 of 64 have the predicted sign. In general the wage and income coefficients are larger for women than for men. This was expected since women have better nonmarket opportunities than men.[8]

### III. The Wage Change Process

We postulate that the rate of change of wages of a particular demographic group is determined by a group specific Phillips curve which gives a role to both wage change expectations and excess demand. Demographic stratification into. sixteen age-race-sex groups captures differences in skills, market opportunities, and the process of wage change which are correlated with demographic characteristics. Since we are considering wage determination within demographic groups, some attention must be given to the transmission of demand and supply across labor market sectors. The rate of wages for a particular group will be affected not only by excess demand in the group's labor market but by conditions in other labor markets as well. For example, if demand for white male labor is high, white male wages rise until it becomes profitable for employers to substitute other types of labor. This substitution process suggests that the rate of wage change in a particular sector is dependent on excess labor demand in all of the sectors. The magnitude of the response of wages in one sector to market conditions in other sectors depends on both the extent of substitution possibilities between types of labor and the relative attractiveness of the jobs in the other sectors. On the demand side, the lack of substitution possibilities among types of labor creates frictions which intensify inflationary wage pressures. On the supply side, labor tends to remain in the sectors with the most attractive jobs, thus creating frictions which limit the mobility of labor into the sectors with poor jobs. This also intensifies inflationary

[8] Because the model is linear and aggregate wages were used, some caution must be exercised in making inferences about the responsiveness of the transitions to demographic wages. Because the aggregate wage overstates female wages and understates male wages, one would expect the coefficient on the aggregate wage to overstate the effect of demographic wages for males and to understate it for females.

pressures. These considerations suggest that the cross-sector effects on wage change are greatest when labor is easily substitutable on the demand side and when jobs are comparable in desirability on the supply side.

The role of wage change expectations in the Phillips curve is based on the notion that employers set wages over a contract period. In setting wages, firms are influenced by both the expected change in the market wage over the contract period and market excess demand. In a multisector framework the rate of change in wages will depend on excess demands in each sector and the expected change in the market wage. If we denote the excess demand for labor in sector $k$ in the year t by $X_{kt}$, and the expected rate of wage change by $\overset{e}{w}_{kt}$, then the rate of wage change in that sector ($w_{kt}$) is determined by a relation such as:

$$(11) \qquad w_{kt} = \sum_j f_{kj}(X_{jt}) + \alpha_k \overset{e}{w}_{kt}$$

where $f_{kj}(X_{jt})$ is the function relating excess demand in sector $j$ to the rate of wage change in sector $k$.

If wage expectations are adaptive according to:

$$(12) \qquad \overset{e}{w}_{kt} = \lambda w_{kt-1} + (1-\lambda)\overset{e}{w}_{kt-1}$$

then it follows that

$$(13) \qquad \overset{e}{w}_{kt} = \frac{\lambda w_{kt-1}}{1-(1-\lambda)D}$$

where $D$ is a delay operator such that $D^n w_{kt} = w_{kt-n}$. Substituting (13) into (11) gives us:

$$(14) \quad w_{kt} = \sum_j f_{kj}(X_{jt}) + \frac{\alpha_k \lambda w_{kt-1}}{1-(1-\lambda)D}$$

which expands to:

$$(15) \quad \begin{aligned} w_{kt} &= \sum_j f_{kj}(X_{jt}) \\ &\quad - (1-\lambda)\sum_j f_{kj}(X_{jt-1}) \\ &\quad + (1-\lambda+\alpha\lambda)w_{kt-1} \end{aligned}$$

If instead of adaptive expectations we assume the expected rate of wage change is proportional to the wage change in the previous period

$$(16) \qquad \overset{e}{w}_{kt} = \lambda w_{kt-1}$$

we get the simpler model:

$$(17) \qquad w_{kt} = \sum_j f_{kj}(X_{jt}) + \alpha_k \lambda w_{kt-1}$$

### *Data and Empirical Results*

Our data analysis using models such as (15) and (17) is still in the exploratory stage. We have constructed an annual time series of earnings by age, race, and sex using both the March Current Population Survey and the Michigan Income Dynamics panel.The Michigan data were used only to derive the distribution of earnings across age groups within race-sex groups. These distributions were then imposed on the *CPS* earnings data for race-sex groups. Preliminary estimates using these data are not encouraging and are not reported here.[9]

## IV. In Conclusion

The dynamic labor force participation relations discussed in this paper are supported by data on short-term participation flows from the Current Population Survey. Our research on the impact of labor markets on relative wage rates is still in an exploratory stage and has yet to reveal any significant labor market effects. While we are unwilling to accept the conclusion that labor market conditions do not affect demographic wage rates, more research will be needed to determine the magnitude of these effects for different population groups.

The effect of wages and earnings on other transition probabilities needs to be

[9] The results are reported in Toikka, William J. Scanlon, and Holt.

explored, especially the quit relation. Since some structural policies affect worker and job availability and others, relative wages and prices, it is encouraging to see a structural model emerging that contains both allocators as a means of influencing workers and employers.

Research in the area of labor market dynamics is relevant for understanding the critical inflation-unemployment problem and policies for dealing with it. Allocation of more resources to basic research in labor markets and inflation will be necessary if we are to deal intelligently with the policy problems before us.

## REFERENCES


F. T. **Denton** *et al.*, (1975a), "Patterns of Unemployment Behavior in Canada," disc. pap. 36, Economic Council of Canada, Ottawa, Sept. 1975.

———, (1975b), "Stock-Flow Relationships and Short-run Dynamics: A Study of the Canadian Labour Market," disc. pap. 37, Economic Council of Canada, Ottawa, Sept. 1975.

C. C. **Holt** *et al.*, *The Unemployment-Inflation Dilemma: A Manpower Solution*, The Urban Institute, Washington 1971.

G. L. **Perry**, "Changing Labor Markets and Inflation," *Brookings Papers*, 1970, *3*, 411–41.

R. E. **Smith**, "A Simulation Model of the Demographic Composition of Employment, and Labor Force Participation: Status Report," working pap. 350–65, The Urban Institute, Washington, July 1974.

——— **and J. E. Vanski**, "The Jobless Rate: Another Dimension of the Employment Picture," working pap. 350–76, The Urban Institute, Washington, May 1975.

———, ——— **and C. C. Holt**, "The Recession Outlook for the Employment of Demographic Groups," *Brookings Papers*, 1974, *3*, 737–60.

R. S. **Toikka**, "A Markovian Model of Labor Market Decisions by Workers," working pap. 350–74, The Urban Institute, Washington, revised August 1975.

———, **W. J. Scanlon and C. C. Holt**, "Extensions of a Structural Model of the Demographic Labor Market," working pap. 350–78, The Urban Institute, Washington, Dec. 1975.

# On the Stability of the Racial Unemployment Differential

*By* ROBERT J. FLANAGAN*

During the post-World War II period, the simple ratio of black-white unemployment rates has shown remarkable stability at about 2:1, despite wide fluctuations in the rate of excess demand in labor markets. As a matter of arithmetic, this implies that recorded black unemployment changes twice as much as white unemployment in response to a given change in aggregate demand. In most discussions, the differential cyclical behavior of black and white unemployment is presumed to be associated with different exposure to layoffs and variously attributed to the tendency of blacks to have less seniority, to receive less specific training, and to be disproportionately employed in cyclically sensitive industries and occupations. Each of these is plausible, although evidence is generally offered only on the third.

While these "facts" may rationalize the observation that the black unemployment rate exceeds and is more cyclically sensitive than the white rate, they do not rationalize the *stability* of the relation until it is explained why the racial differences in job assignment, seniority, and specific training remain stable over time. A common response is that they reflect a stable intensity of discrimination over the period. But this merely restates the question, for there has been a substantial increase in the amount of private and public antidiscrimination activity since the early 1960's, and this should have had some impact on patterns of occupational assignment, seniority, and specific training of blacks.

The theme of this paper is that the stability of the aggregate racial unemployment ratio obscures important recent changes in the relative unemployment exposure of blacks in demographic subgroups defined by schooling, marital status, age, and labor force experience. Moreover, the key changes are most clearly revealed by separating unemployment experience over the life cycle (i.e., changes in unemployment associated with the aging or increased experience of a labor force cohort) from vintage effects (changes in the unemployment exposure of successive labor force cohorts) and by deemphasizing explanations based on racial differences in layoffs. The analysis of vintage and life-cycle changes in black relative unemployment also permits preliminary tests between competing explanations of racial unemployment differentials and changes in those differentials. In particular, it offers a solution to the puzzle of why the narrowing of racial earnings differentials in the late sixties was not accompanied by as dramatic and sustained a narrowing in unemployment differentials.

## I. Life Cycle and Vintage Hypotheses

### *A. Search Behavior*

Even in a color-blind labor market, unemployment would vary over the life cycle, for job search investments are rationally concentrated in early years of labor force

* Associate Professor, Graduate School of Business, Stanford University. Work on this paper was supported by contract no. L–74–131 from the Office of the Assistant Secretary for Policy Evaluation and Research in the U.S. Department of Labor. The paper does not necessarily represent the views of the Department of Labor, its agencies or staff.

experience. Moreover, the relative unemployment exposure of younger workers is further increased if the appropriate wage distribution is not known in advance and if there are systematic tendencies toward over-optimistic wage expectations among youth. Early periods of unemployment then serve to bring wage expectations into line with wage offers.

A racial unemployment differential can emerge if blacks expect to receive offers from the same wage distribution as whites but in fact face a lower distribution of wage offers. With labor force experience, the wage expectations of both races should become more accurate and their unemployment rates should decline. Moreover, as blacks learn that the distribution of wage offers differs by race, and their wage expectations adjust accordingly, search-motivated unemployment exposure of blacks and whites should become more similar. This important life-cycle implication of a converging race differential distinguishes the unemployment predictions associated with search behavior from some other explanations. In general, however, the convergence will stop short of equality of the black and white unemployment rates to the extent that a net racial wage differential persists, for this will tend to raise the quit rate and the time-intensity of search by blacks.

### *B. Labor Quality*

Measures of both the inputs (resources) and outputs (completion and achievement score levels) of the schooling system indicate that (1) the quality of each year of schooling is lower for blacks than whites and (2) the racial difference in school quality has been narrowing over time (Finis Welch 1973). In principle the school quality differentials could have both direct and indirect influences on racial unemployment differences.

To the extent that lower school quality affects job performance, workers who receive lower quality schooling may be systematically assigned to jobs which are relatively sensitive to seasonal and cyclical fluctuations and, in general, might accrue less seniority and protection from layoff. However, racial unemployment differentials which result from the *direct* effects of school quality differentials should be relatively *stable* over the life cycle after completion of schooling.

On the other hand, if relatively low school quality limits (or is believed to limit) an individual's ability to benefit from specific training, racial school quality differentials may have further indirect effects on unemployment. In particular, if specific training is a positive function of school quality, and if specific training accumulates with the aging (labor force experience) of a cohort, then racial differences in quits, layoffs, and associated unemployment exposure should *widen* over the life cycle as whites received relatively large specific training investments. Both the potential direct and indirect effects of racial school quality differences imply a life-cycle pattern of racial unemployment differentials that differs sharply from the narrowing life-cycle differentials implied by the search behavior explanation and which can be examined in the data below.

In contrast to the search behavior story, the labor quality explanation yields clear implications concerning vintage effects. If the *relative* quality of black schooling has been increasing over time, as evidence seems to indicate, the racial unemployment differential attributable to quality differences should be smaller in each successive cohort for a given length of labor force experience.

### *C. Declining Wage Discrimination*

Equal opportunity policies of the past decade have had two major elements, both of which should influence the pattern of

unemployment among blacks and whites. One objective, associated with affirmative action programs, has been to increase the relative employment and utilization of labor force minorities and in the case of blacks has apparently stimulated demand most for new cohorts of college graduates (see below). The other objective has been to achieve something approximating equal pay for equal work and has had a more complex impact.

Theories of discrimination based on aversion toward blacks or methods of screening workers of uncertain quality predict that institutional constraints on racial wage discrimination will tend to widen racial unemployment differentials. In the case of employer aversion, racial wage differentials compensate for prejudice; in the case of statistical discrimination, they compensate risk-averse (but unprejudiced) employers for risk when the testing process is less reliable for blacks than whites. With full wage compensation, theories predict no racial differences in the probabilities of layoff or hire. However, employers will adjust to equal pay for equal work laws and other constraints on wage discrimination by practicing discrimination in employment.

Nevertheless, these propositions can provide an erroneous guide to the interpretation of aggregate unemployment differentials for two reasons. First, accepting a lower wage does not empirically reduce the probability that a black will be laid off. Second, the racial wage difference tends to increase the relative quit rate and the time intensity of search among blacks. (For evidence, see Flanagan, 1975.) Therefore, the net effect of a reduction in racial wage discrimination should be to *reduce* racial unemployment differentials among *experienced* workers as black quits and the time intensity of black search fall.

The effect on experienced workers is stressed, because a fall in wage discrimination will raise the value of market relative to nonmarket activity for blacks and tend to increase the flow of black labor force entrants which will normally increase unemployment during the transition. The general relationship between changes in racial wage and unemployment differentials will depend on the relative magnitude of the offsetting unemployment adjustments among experienced and inexperienced workers. In large measure this will simply reflect the proportion of a particular age or experience cohort which is not in the labor force. Clearly the size of the potential entrant flow is largest in the youngest cohorts where the percent out of the labor force is relatively large. On the other hand, it is unlikely that unemployment associated with induced labor force entry would be sufficiently large to overwhelm the falling unemployment differential for experienced workers among prime-age males.

The vintage effects of falling wage discrimination are now clear. Each successive cohort of experienced workers should have a smaller racial unemployment differential. The vintage effect may be difficult to observe in the youngest cohorts because of the difficulty of isolating experienced and inexperienced workers in the data (except by age). Indeed, it is possible that racial differences may widen in groups dominated by inexperienced workers. An empirical analysis of these effects follows.

## II. Empirical Analysis

### *A. Life-Cycle Developments*

Table 1 is constructed to reveal changes in the relative unemployment of black males associated with aging ten years (1960–1970), after schooling, marital status and age cohort are held constant. For convenience, most of the discussion is focused on men who are married with

TABLE 1—WITHIN-COHORT CHANGES IN MALE BLACK-WHITE UNEMPLOYMENT RATE RATIO, ALL SCHOOLING GROUPS, 1960-70

| Age in 1970 | Married, Wife Present | Other |
|---|---|---|
| 25+ | −.43 | −.19 |
| 25-29 | .17 | .06 |
| 30-34 | −.05 | −.27 |
| 35-44 | −.54 | −.33 |
| 45-54 | −.74 | −.38 |
| 55-64 | −.56 | −.26 |

*Source:* U.S. Bureau of the Census, 1963, 1973.

wife present. For the decade of the 1960's, the data show a substantial, across-the-board decline in the relative unemployment of blacks twenty-five years of age and over in 1970 (16+ in 1960). Moreover, the falling relative unemployment exposure of experienced blacks is observed in virtually all age cohorts, with the largest declines associated with the aging of the older cohorts, in which the unemployment ratio was greatest in 1960.

These data do not support the notion that racial differences in school quality and specific training were the dominant influence on life-cycle racial unemployment patterns for the period 1960–70. It is equally doubtful that search motivations are the sole explanation for the life-cycle decline in the relative unemployment of blacks. If search motivations were the dominant influence on racial unemployment patterns, one would expect that the life-cycle declines would be smaller in the older cohorts. Since the search explanation rests on the assumption that blacks *initially* fail to distinguish between the full market wage distribution (for a given skill) and the distribution of wage offers to blacks, racial unemployment differentials should converge most rapidly as blacks learn that the distributions differ. But this is a process which should be completed within a few years of labor force experience. At best, this pattern is observed only weakly in the data.

*Schooling.* In 1960, the racial unemployment ratio varied substantially by age cohort and schooling level. In general, the relative unemployment exposure of blacks was greatest at the highest schooling levels, where the black labor force share was disproportionately low. During the 1960's racial schooling differences continued to narrow, increasing the relative supply of blacks in higher schooling categories. In the absence of shifts in the relative demand for blacks or a fall in the

TABLE 2—LIFE-CYCLE CHANGES IN RELATIVE BLACK UNEMPLOYMENT BY YEARS OF SCHOOLING, 1960–70
(Men married with wife present)

| Age in 1970 \ Schooling | All | 8 | 9–11 | 12 | 13–15 | 16 |
|---|---|---|---|---|---|---|
| 25+ | −.43 | −.54 | −.58 | −.52 | − .84 | −1.78 |
| 25–29 | .17 | −.63 | −.06 | −.32 | −1.56 | — |
| 30–34 | −.05 | .03 | −.34 | −.37 | .06 | −3.74 |
| 35-44 | −.54 | −.43 | −.61 | −.43 | −1.38 | −1.28 |
| 45–54 | −.74 | −.68 | −.78 | .77 | −1.49 | −1.80 |
| 55-64 | −.56 | −.65 | −.49 | −.66 | − .50 | −2.26 |

*Source:* see Table 1.

black relative wage, the increased relative schooling would have raised the relative unemployment of blacks as the black labor force grew most rapidly in schooling levels which had the highest unemployment exposure in 1960.

But in reviewing the data across schooling categories (see Table 2), it is clear that (1) declines in black relative unemployment are virtually across the board and (2) the largest declines over the decade are in the post-high school groups which recorded the highest relative unemployment in 1960. These declines therefore offset any opposite tendencies attributed to the shifting labor force composition of blacks toward higher schooling levels. By 1970 the unemployment rates of blacks and whites were near equality for those with eight years *and* for those with sixteen years of schooling.

Since the relative wage of highly-schooled blacks increased over the decade (see Richard Freeman, J. P. Smith and Welch), the data are consistent with the view that the employer response to changing demand and growing antidiscrimination activity has been to hire the most educated and the most recently graduated blacks first. For the cohorts over 25 years of age in 1960 (35 and over in 1970), an increase in the relative supply of schooling was generally not feasible. Given the inelastic supply of higher schooling in these age cohorts, increases in employer demand achieved substantial reductions in relative unemployment (and increased relative earnings). In younger cohorts, schooling was not necessarily complete in 1960, and individuals could therefore adjust their school-leaving plans in response to shifts in demand. The supply of labor with a given level of schooling should be more elastic in these cohorts, implying a somewhat smaller decline in relative unemployment (increase in relative annual earnings) if increases in demand were spread evenly over all cohorts. In fact, the largest decline in relative unemployment was for men aged 20–24 in 1960 who had completed 16 years of schooling by 1970. This suggests that increases in relative demand for black workers were greatest for recent college graduates.

TABLE 3—CHANGE IN RACIAL UNEMPLOYMENT RATE RATIO, 1960–70

| Age | Married, Wife Present | Other |
|---|---|---|
| 16+ | −.38 | −.05 |
| 16–24 | 0 | .18 |
| 16–19 | .18 | .28 |
| 20–24 | −.03 | .10 |
| 25+ | −.45 | −.22 |

*Source:* see Table 1.

### *B. Vintage Effects*

The pattern of vintage effects is summarized in Table 3. The prevalence of negative signs indicates that in general the more recent black cohorts at each age have lower relative exposure to unemployment than a decade earlier. However, this generalization is somewhat qualified along the dimensions of labor force experience and marital status. The greatest narrowing of unemployment differentials occurred for more experienced cohorts (25 years and older) who had largely completed their schooling by 1960. Moreover, strong vintage reductions in racial unemployment differentials also occurred within schooling groups. (The vintage reductions in *relative* unemployment were largest for workers who were married with wife present. Because unemployment rates are generally lower in the "wife present" marital status category, however, the most extensive narrowing in racial unemployment *differences* occurred in the "other" category in which the racial differences were greatest at the beginning of the period.)

However, the vintage effects among in-

experienced workers are in the opposite direction. Over the decade, the racial unemployment differential increased in the 16–19 year group and was focused on the "other marital status" group, which comprises the majority of the labor force at this age.

This particular pattern of vintage effects is not fully consistent with the labor quality view, which does not explain why racial unemployment differences have declined for new vintages of experienced workers but increased among new vintages of teenage workers. As noted above, however, the pattern of vintage effects can be explained by a fall in wage discrimination and generally improving market opportunities for blacks. In age groups where the potential size of the induced flow is large (that is, groups with relatively low labor force participation rates), the tendency of racial unemployment differences to fall among experienced workers in response to greater equality of opportunity may be offset by rising unemployment among new labor force entrants. The crucial link in this behavioral explanation of the vintage effects is a rise among teenage blacks of unemployment associated with labor force entry.

This issue was pursued using quarterly *BLS* data on unemployment by reason, which have been available in reliable form only since 1968. For the period 1968–73, there was a trend increase in black (but not white) unemployment after adjustment for cyclical and seasonal influences. It was possible to decompose the black trend into the effects of trends in unemployment associated with layoffs, quits, labor force entry and reentry among adults and teenagers as well as changes in the demographic composition of the black labor force. The decomposition of the 1968–73 trends in the unemployment rates by source indicates that the main contributory factors were statistically significant secular increases in the new entrant rate among black teenagers of both sexes. (Unreported regression results are available from author.)

At the root of the "puzzle" of the stability of the aggregate racial unemployment ratio during a period in which other indicators of racial differences in economic status narrowed considerably is a failure to distinguish between the impact of antidiscrimination programs on experienced and inexperienced workers. Among groups with a high labor force participation rate, the dominant influence should be a narrowing of unemployment associated with quits and layoffs, and in fact the Census data show that race differentials have narrowed substantially among experienced workers. However, one effect of antidiscrimination programs can be to raise expectations of the level of market relative to nonmarket rewards. This will tend to raise the rate of gross labor force entry among groups in which many individuals previously opted for nonmarket activity. Youth is the most obvious example, and we observe that (1) the relative unemployment of black males with little work experience has increased while the relative unemployment of experienced black workers had declined, and (2) virtually all of the recent trend *increase* in the unemployment of teenage black males is associated with labor force entry. Alternative explanations, most notably the influence of minimum wage legislation, fail to account for why the trend increase appears to be restricted to black youth during a period in which racial quality differentials have narrowed substantially. Moreover, over the 1968–73 period, the ratio of the basic minimum wage to private nonfarm average hourly earnings (adjusted for coverage) declined, so that the influence of minimum wage legislation on general teenage unemployment should have dissipated.

REFERENCES

**R. J. Flanagan,** "Discrimination Theory, Labor Turnover, and Racial Unemployment Differentials," mimeo, Apr. 1975.

**R. Freeman,** "Decline of Labor Market Discrimination and Economic Analysis," *Amer. Econ. Rev. Proc.*, May 1973, *63*, 280–86.

**J. P. Smith and F. R. Welch,** "Black-White Earnings and Employment: 1960–1970," Rand Corporation, mimeo, 1975.

**F. Welch,** "Black-White Differences in Returns to Schooling," *Amer. Econ. Rev.*, Dec. 1973, *63*, 893–907.

**U.S. Bureau of the Census,** *Census of Population, 1960*, Subject Reports, Final Report PC(2)–5B, *Educational Attainment*, Washington 1963.

———, *Census of Population, 1970*, Subject Reports, Final Report PC(2)–5B, *Educational Attainment*, Washington 1973.

# Occupational Mobility and the Distribution of Occupational Success Among Young Men

By Robert E. Hall and Richard A. Kasten*

Interpretation of the distribution of occupational success depends crucially on an understanding of the process of mobility. If the holders of low-paying jobs this year are just as likely as anyone else to hold good jobs next year, then the lower tail of the earnings distribution is not a matter of great social concern. Further, if the distribution of success is primarily determined by a random process, the prospects for a significant reduction in the inequality of earnings are unfavorable. On the other hand, high mobility in the short run may conceal systematic stratification of the labor force—it may be that the same individuals spend most of their time in the worst jobs. If so, luck and random events may play a small role in the distribution of well being, inequality may be a major issue, and the elimination of stratification may hold out the promise of an important reduction in inequality.

Simple tabulation of distributions of success and of mobility among categories of success cannot answer the basic question of the role of stratification. In this paper, we develop a model and statistical method for studying mobility and relating it to characteristics upon which stratification may be based. Our study rests on annual interviews conducted from 1966 through 1969 with about 3,000 white males born between 1942 and 1952.[1] We find that mobility is closely related to personal characteristics, especially intelligence and education. Men with high *IQ*s and extensive education are more likely to move to high-paying jobs and are more likely to remain in them. Nevertheless, random events dominate the distribution of occupational success in the short run. Our results strongly support the finding of Christopher Jencks and his collaborators that redistribution of the determinants of earnings would do almost nothing to reduce the observed inequality of annual earnings.

## I. The Model

Our concern is with the distribution of occupational success, not just with its expectation, so we are obligated to formulate a model that is explicitly probabilistic. Our first step is to make occupational success a discrete measure taking on only five alternative values (construction of the measure is described in the next section). The distribution of success is the fraction of the labor force in each of the five categories. The process of occupational mobility is portrayed within the model in a set of three probability distributions. The first is the simple binary probability that a worker will change jobs in a given month; this probability depends on his age, intelligence, and education. The second is the five-way distribution of the category of the new job for job-changers, which depends on personal characteristics and on the category of the previous job. The third probability applies to workers

* Dept. of Economics, Massachusetts Institute of Technology and the Dept. of Health, Education and Welfare, respectively.

[1] In this paper we concentrate on whites. In an earlier paper (see list of references), we studied black-white differences in the same framework.

who have remained on the same job for at least 12 months and gives the distribution among job categories as the result of promotion or demotion within the same firm. Again, this distribution depends on personal characteristics and on the previous occupational category. The model also treats education as an endogenous variable. Each year, there is a probability that an individual will return to school in September. This probability depends on personal characteristics, especially *IQ*, and is much higher for those individuals who were in school the previous year. For those who leave school and enter the labor market, there is a probability distribution for the occupational category of the first job. The five major probabilities of the model will be called the "change job," "new job," "promotion," "return to school," and "first job" equations.

The model is sufficiently general to encompass the two extreme views that mobility is a purely random process, on the one hand, and that stratification is absolute, on the other hand. Pure randomness would reveal itself as identical probability distributions for all individuals, independent of intelligence, family background, or other immutable characteristics. Absolute stratification would show up as probability distributions that made unambiguous predictions of different careers for differently endowed individuals, making the observed distribution of mobility entirely the result of the distribution of endowments. Between the two extremes, the more sensitive are the probabilities to the personal characteristics, the more evidence there is of stratification.[2]

## II. Data

Our data are taken from the National Longitudinal Survey of Work Experience (the Parnes data) for young men. From the annual surveys, we reconstructed a monthly chronology of activities of the members of the sample.[3] We created the five occupational categories by the following procedure. First we imputed earnings to each 3-digit occupation as the average annual earnings of individuals holding those occupations for 50 or more weeks in 1959. We then defined five categories, chosen so that our sample was roughly evenly distributed among them. These are

category 5: more than \$5900
category 4: \$5100 to \$5900
category 3: \$4600 to \$5100
category 2: \$3700 to \$4600
category 1: \$3700 or less.

After detailed study, we found this approach to occupational classification superior to its two main competitors: the Duncan-Blau socioeconomic score (*SES*),

[2] Econometrically, the dependence of the probability on personal characteristics and other determinants is indexed within our model by the parameters of a logit function:

$$p_i(x_t) = \frac{e^{x_t \beta^i}}{\sum_{j=1}^{N} e^{x_t \beta^j}}$$

$p_i\,(x_t)$ is the probability of the $i$th alternative, as follows:

change job:
$i=1$ no change
$i=2$ change
return to school:
$i=1$ return
$i=2$ enter labor market
new job, promotion, first job:
$i=1,\ldots,5$ for 5 occupational categories;

$x_t$ is a vector of immutable personal characteristics (family background and *IQ*), human capital (education and experience measured by time on the job) and market effects (captured by a set of dummy variables for years). Each alternative has a vector of parameters, $\beta^i$, which determine the influence of the variables, $x_t$, on the probability that $i$ will occur. We normalize the $\beta$s by setting the vector $\beta^1$ to zero, so each of the other $\beta^i$s controls the probability of the other alternatives relative to the probability of the first. In presenting the results, we focus on the $p_i$ values implied by our estimates of the $\beta$s rather than on the $\beta$s themselves. Estimation of the $\beta$s was carried out by conventional techniques of maximum likelihood.

[3] An appendix describing the data in more detail is available from the first author.

which attempts to take account of the nonmonetary rewards to occupations, and the actual earnings reported in the survey. The *SES* has a pronounced bias against blue-collar jobs which is particularly inappropriate for our sample. Actual earnings are reported for most but not all of the jobs held by members of the sample. However, among young workers, an important component of the return to labor in some jobs is the accumulation of training on the job (see Jacob Mincer). We concluded that the earnings of workers of all ages holding a given occupation is a better measure of the economic value of the occupation than is the reported wage of a young worker.

## III. Properties of the Individual Components of the Model

### *A. Education*

The typical young man remains in school until the end of the school year and then takes a job at the beginning of the summer. At the end of the summer he faces the central decision whether to remain at work full time or to return to school. The probability of returning to school is the determinant within our model of the amount of education received by each individual and is a function of *IQ*, family background, age, and grade last completed, as well as other variables of lesser importance. The combined influence of age and last grade make the model predict the appropriate distribution of total grades completed, especially the large fraction that completes exactly 12 grades. For the questions addressed by this paper, the important variables here are *IQ* and family background (measured by the father's *SES*). Both have the expected effect—more favorably endowed individuals tend to continue longer in school. *IQ* is the stronger of the two influences. Individuals who are 10 points above the average *IQ* of 100 have a 64 percent probability of going on to college after high school, against 53 percent if *IQ* is 100. By contrast, individuals whose fathers have *SES*s 32 points above the average of 36 have only a 59 percent probability of college, against the same 53 percent if *SES* is 36. The two variables have comparable effects at other stages in the educational process. Since education and the time of entrance into the labor market have profound effects on occupational success, the impact of these two variables on the decision to continue in school is a major stratifying influence.

### *B. First Job*

The occupation of the first job has a central role in our analysis. Some young men remain in their first job throughout the span of our study, and for those who do change occupations, the earlier occupation has a strong influence on the new occupation. The first job marks the transition between the accumulation of general human capital at school and the accumulation of occupation-specific human capital on the job. The latter process is one of the major explanations of the persistent effect of the early occupation on the subsequent career.

It is impossible to separate the effects of age and education on the first job, since at the time most men are six years older than the grade they complete. The combined influence of the two is dramatic and shows clearly that the process that assigns men to first jobs is far from random. High-school dropouts who enter the labor market at age 16 typically find poor jobs—44 percent in the lowest category and another 24 percent in the second category. Only 13 percent find first jobs in the top two categories. At the other extreme, college graduates entering at age 22 go overwhelmingly into the top of the occupational structure—58 percent into category 5 and another 19 percent into category 4.

Only 18 percent of them fall into the two lowest categories. High school graduates are between the two extremes, distributed by category as follows: 1. 33 percent, 2. 24 percent, 3. 17 percent, 4. 18 percent, and 5. 8 percent.

Apart from their indirect effects through education, intelligence and family background make roughly equal direct contributions to the level of the first occupation. For high school graduates, an individual with an *IQ* of 115 has a 30 percent probability of taking a job in category 1 and a 10 percent probability in category 5, against 33 percent and 8 percent respectively for those with the average *IQ* of 100. Individuals whose fathers have an *SES* of 68 had only a 26 percent probability in category 1 and 9 percent in category 5, against 33 percent and 8 percent for those whose fathers have the average *SES* of 36. Not surprisingly, having a well-placed father seems to help most in keeping the son out of the worst jobs.

### *C. Promotion*

In the Parnes data, occupation is reported annually for young men who do not change employers. About one-quarter of the workers change occupations, and of these, two-thirds are promotions (movements into higher occupational categories) and the rest demotions. For high school graduates with average endowments (*IQ* of 100 and father's *SES* of 36), the model implies the matrix among occupational categories shown in Table 1. Particularly noteworthy is the high probability of promotion out of the two lowest categories. Workers who are able to remain with the same employer move rapidly out of the worst jobs. The lower tail of the distribution of workers among occupational categories does not represent workers who are permanently trapped in poor jobs within the same firm. The mobility process for young workers has a pronounced upward bias; almost all the upper right-hand elements of the transition matrix are larger than the corresponding lower left-hand elements.

TABLE 1—MATRIX OF OCCUPATION CATEGORIES FOR HIGH SCHOOL GRADUATES WITH AVERAGE ENDOWMENTS WHO DO NOT CHANGE EMPLOYERS

| Category Last Year | Category this Year | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| 1 | .50 | .18 | .16 | .08 | .07 |
| 2 | .02 | .65 | .09 | .15 | .09 |
| 3 | .01 | .07 | .73 | .11 | .08 |
| 4 | .00 | .08 | .07 | .76 | .08 |
| 5 | .02 | .06 | .07 | .12 | .73 |

Endowments and education have small but important influences on advancement within the firm. High-school graduates previously in category 3 have a 19 percent probability of promotion to categories 4 or 5, while the same probability for workers with two more years of education is 20 percent. A 15-point increment in *IQ* raises the probability for the high school graduate from 19 percent to 23 percent. Father's *SES* has no systematic effect on promotion or demotion. The effects of education and *IQ* cumulate year after year in the process as long as the worker remains with the same firm, so their long-run effects are much larger than is suggested by the small shifts in the transition matrix they induce.

### *D. Job Changes*

About 6 percent of the working members of our sample lose or leave their jobs each month (the data do not permit us to distinguish the two sources of changes). Virtually all of them find new jobs within a month or two, often with an intervening spell of unemployment. The rate of turnover is an important determinant of the dispersion of occupational success, as

TABLE 2—MATRIX OF OCCUPATION CATEGORIES FOR HIGH SCHOOL GRADUATES WITH AVERAGE ENDOWMENTS WHO CHANGE JOBS

| Category of Old Job | Category of New Job | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| 1 | .19 | .26 | .21 | .21 | .13 |
| 2 | .07 | .46 | .14 | .22 | .11 |
| 3 | .06 | .18 | .47 | .18 | .11 |
| 4 | .05 | .17 | .16 | .52 | .09 |
| 5 | .06 | .20 | .14 | .18 | .43 |

large upward and downward occupational movements are much more likely to occur when workers move between employers. Endowments and education seem to have almost no direct effect on the turnover rate. Turnover is highest among young workers, falling from 9 percent per month at age 16 to 4 percent per month at age 24. Workers are much more likely to leave or lose jobs in the lowest category (8 percent per month) than in the highest category (5 percent per month); the percent falls smoothly between the two.

### E. *New Jobs*

Most job changers improve their occupational category, or at least remain in the same category. The transition matrix for changers is shown in Table 2. Again, it is noteworthy that few workers remain in the worst jobs—those in the lowest category have both the highest probability of changing jobs and the highest probability of moving upward as a result of the change. The general bias of mobility through job changes is upward, but not as strongly as in the case of movements within the firm. Movements through the labor market are clearly riskier than movements within the firm.

Endowments and education have substantially more effect on the distribution of job changers among occupational categories than in the promotion equation. A high-school graduate has a probability of 20 percent of advancing from a category 3 job to categories 4 or 5, but this probability rises to 36 percent for an otherwise identical worker with 14 years of education. Fifteen extra points of *IQ* raises the same probability from 29 percent to 33 percent, and 32 extra points of father's *SES* raises it from 29 percent to 31 percent. Since job changes frequently involve dissipation of occupation- and firm-specific human capital, it is not surprising that the general determinants of success matter more in this equation and the category of the previous job matters less, relative to the promotion equation.

## IV. Properties of the Complete Model

When the probabilities described in the previous section are permitted to interact with one another, they imply a distribution of individuals across occupational categories at any given time. Within a group that is homogeneous with respect to

TABLE 3—DISTRIBUTION AMONG JOB CATEGORIES BY IQ, FATHER'S *SES* AND EDUCATION

| | Job Category | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| *IQ* (full effects) | | | | | |
| 90 | .11 | .14 | .22 | .26 | .28 |
| 100 | .10 | .13 | .17 | .25 | .35 |
| 110 | .10 | .11 | .13 | .25 | .42 |
| *IQ* (holding grades at 12) | | | | | |
| 100 | .11 | .14 | .18 | .27 | .29 |
| 110 | .10 | .14 | .15 | .28 | .33 |
| Father's *SES* (full effects) | | | | | |
| 16 | .13 | .13 | .18 | .27 | .30 |
| 36 | .10 | .13 | .17 | .25 | .35 |
| 60 | .08 | .12 | .18 | .22 | .41 |
| Grades completed | | | | | |
| 10 | .11 | .15 | .24 | .30 | .20 |
| 12 | .11 | .14 | .18 | .27 | .29 |
| 14 | .10 | .12 | .14 | .21 | .43 |

the fundamental determinants of success (*IQ* and family background), those who are in the lowest category are the victims of an unfavorable random event. Some have just passed through the labor market and have taken a poor job; if they remain with their new employers, they are likely to move to better jobs fairly soon. Others may have left school unexpectedly early, in which case the model predicts that they are more likely to remain in the lower categories throughout their careers. The distributions of workers among categories induced by the complete model are shown in Table 3. The first three rows, labeled "*IQ* (full effects)," shows the distribution among occupations of a group of 24 year old men whose fathers all had the average *SES*, 36. All three distributions have a great deal of dispersion. Both the short-run randomness embodied in the mobility equations of the model and the long-run randomness in the educational process contribute to this dispersion. Bright individuals are almost as likely to be found in the worst jobs as a result of bad luck as are those with below average intelligence. On the other hand, a higher *IQ* substantially improves a worker's chances of landing a top job. Underlying the randomness in the assignment of workers to occupations is an important tendency favoring the more intelligent.

The next two rows of Table 3 show that most of the systematic effect of *IQ* operates through the educational process. When the probability distribution is computed conditional on exactly 12 years of school, the upward shift in the distribution for brighter individuals is much weaker. The next part of the table shows the full effects of family background, measured by the father's *SES*. The sons of well-placed fathers are at a substantial advantage in the labor market, though they are subject to the same random processes as everyone else. They are both less likely to be at the bottom of the occupational structure (this is the influence of the first job equation) and more likely to be at the top, when compared to otherwise identical men with less successful fathers. Finally, the last three rows of Table 3 show the major role of education in stratification at the upper occupational levels. Within the group of workers with average *IQ* and father's *SES*, those who obtain more education are much more likely to be in the best jobs. However, well-educated men face almost the same probability of holding the worst kind of job as do high school dropouts.

Table 4—Estimated Effects of Redistributing Individual Characteristics on the Distribution Among Job Categories

| | Job Categories | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| Characteristics distributed as in sample | .12 | .13 | .18 | .22 | .35 |
| *IQ* held constant at 100 | .11 | .14 | .16 | .26 | .33 |
| Father's *SES* held constant at 36 | .10 | .13 | .18 | .22 | .37 |
| Grades completed held constant at 12 | .11 | .14 | .20 | .24 | .31 |
| All three held constant | .11 | .14 | .18 | .27 | .29 |

Table 3 shows that the determinants of occupational success matter a great deal. Men with high *IQ*s, with successful fathers, and with college educations are very clearly separated from those with less favorable backgrounds by their much higher probability of holding top jobs. The hypothesis that the underlying determinants of success are unimportant compared to the random influences is refuted by these results. Nonetheless, it is true that redistribution of the determinants of success would do almost nothing to reduce the unequal distribution of success. This controversial proposition, which has been advocated by Christopher Jencks

and his associates, is illustrated in Table 4.[4] In the first line, we show the distribution of the members of the sample predicted by the model, using the actual dates of birth, *IQ*s, and fathers' *SES*s (this distribution matches the actual distribution of the sample quite closely). In the second line, we show the hypothetical distribution within the sample if all its members had the average *IQ*. The next two rows repeat the exercise for father's *SES* and years of education separately, and the last line considers the result of making the sample homogeneous with respect to all characteristics except age. In all cases the results are striking. The model predicts that 25 percent of the sample will hold jobs in the two lowest categories. The only redistributional experiment that reduces this figure at all is the one that makes fathers' *SES* uniform. The only noticeable impact of redistribution is to shift workers from the fifth to the fourth categories.

[4] Our model describes the labor market only from the point of view of the worker, and does not deal with changes in wages caused by changes in supply. Our results on the impact of redistribution of endowments are strictly applicable only if the job categories are close substitutes on the demand side. Any lack of substitutability among the categories would make the impact of redistribution even smaller. Our results give an upper bound to the potential effect of redistribution.

Our findings can be summarized in two simple propositions:

(1) Intelligence, family background, and education matter a great deal. The occupational success of the entire labor force could be improved materially by increasing its endowments of any or all of the three.

(2) Redistribution of endowments would have almost no effect on the distribution of success. The favorable effects for the fraction of the labor force with below-average endowments would be largely offset by the unfavorable effects on those above the average.

## REFERENCES

**R. E. Hall and R. A. Kasten,** "The Relative Occupational Success of Blacks and Whites," *Brookings Papers*, 1973, *3*, 781–95.

**C. Jencks,** *Inequality: A Reassessment of the Effect of Family and Schooling in America,* New York 1972.

**J. Mincer,** *Schooling, Experience, and Earnings,* New York 1974.

# The Entrepreneurship Decision and Black Economic Development

*By* William D. Bradford and Alfred E. Osborne, Jr.*

In the last few years there have been tremendous efforts by various government agencies to promote minority entrepreneurship. These programs, the most notable of which was President Nixon's "Operation Mainstream," were intended to aid the process of black economic development through increased minority business formation. Most programs, however, appeared to simply encourage employed blacks to discontinue their salaried or wage occupation and attempt to become independent businessmen.

The extent to which black entrepreneurs can be mobilized to economically develop the community is not clear. If by economic development one primarily refers to sustained increases in per capita income of a given community and accepts that as the relevant definition,[1] it is important to understand how a policy which encourages primarily salaried individuals to become entrepreneurs might affect the relative standard of living in minority communities. This paper explores some basic propositions implied by such decisions.

## I. The Entrepreneurship Decision

The entrepreneurship decision can be considered a specific case of the many occupational choices available to an individual. Self-employment, however, has not received the attention given wage and salaried employment. The intensive analysis recently given the relationship between income, education, and occupational choice has virtually neglected those decisions which might result in a supply of entrepreneurs.[2]

While there are a few studies available which report on returns to the entrepreneurial sector, most are old, inconclusive, and often based on poor data.[3] Lauritis R. Christensen, however, reports that entrepreneurial labor and capital earn a rate of return approximately equal to that earned by corporate factors. These results are contrary to previous studies which report that entrepreneurial returns are smaller. While much more work needs to be done in estimating returns, the recent work does suggest that self-employment is a competitive alternative to business employment even though as a percent of national income entrepreneurial activities are on the decline.

Osborne (1975) has characterized the entrepreneurship decision as involving the choice between two income streams. In that paper a rational individual was assumed either (1) to accept (continue in) a wage position earning a certain amount of dollars per year or (2) to go into business with the anticipation of realizing

* The authors are Assistant Professor of Finance, Stanford University and Assistant Professor of Business Economics, University of California, Los Angeles, respectively. Helpful comments from colleagues Thomas Copeland and David Mayers are acknowledged.

[1] See for example Gerald M. Meier.

[2] One example is W. Lee Hansen.

[3] Previous studies include Stanley Lebergott and D. Gale Johnson among others.

some share of the economic profits. It was formally shown that if T is the projected length of business involvement and $Y_t$ the yearly wage to be foregone, then the decision point comparison of total returns from business activity consisting of periodic withdrawals and other payments ($R_t$) plus terminal value ($V_T$) to employment can be expressed as[4]

$$(1) \qquad \sum_{t=0}^{T} \frac{Y_t}{(1+r_t)^t} \leq \sum_{t=0}^{T} \frac{R_t}{(1+\rho_t)^t} + \frac{\chi \cdot V_T}{(1+\rho_T)^T}$$

*where* $\chi$ is his share of the terminal equity value of the firm ($0 < \chi \leq 1$), $r_t$ is the rate at which wage or labor income is discounted and $\rho_t$ is the rate at which income from business activities is discounted in period t. For convenience, we will assume throughout our analysis that $\chi = 1$. This amount of wage income (in present value terms) would be willingly exchanged for entrepreneural income if the latter income stream is greater. A potential entrepreneur, black or white, is therefore assumed to base his decision on $V_T$, $Y_t$, $R_t$, $\rho$, $r$, and $T$. Further, any such entrepreneur is indifferent if only the weak equality condition in equation (1) holds.

Let us consider two representative entrepreneurs, one black ($B$) and one white ($W$). Define the left-hand side of equation (1) as $\hat{Y}_W$ and $\hat{Y}_B$; the two terms on the right-hand side as $\hat{R}_W + \hat{V}_W$ and $\hat{R}_B + \hat{V}_B$ such that

$$(2) \qquad \hat{Y}_B \leq \hat{R}_B + \hat{V}_B \quad \text{and} \quad \hat{Y}_W \leq \hat{R}_W + \hat{V}_W$$

[4] The value of human capital in period $T$, which will be positive under either alternative, is not herein analyzed. First, if $T$ is large, then the value of human capital will be small. Secondly, since the theory has not yet resolved the issue of the relative value of employment and entrepreneurship human capital (see Christensen, pp. 584–85), a simplifying assumption is that employment and entrepreneurship human capital in period $T$ are equal, and thus cancel each other out. In addition, external effects and nonpecuniary reasons for making this choice are not directly considered.

We can also extend the model to consider investment capital inputs (equity) by entrepreneurs which represent a portion of their accumulated wealth or savings. If $W_B$ *and* $W_W$ represent the initial capital input of a black and white entrepreneur respectively, expression (2) becomes

$$(3) \qquad \hat{Y}_B + W_B \leq \hat{R}_B + \hat{V}_B \quad \text{and} \quad \hat{Y}_W + W_W \leq \hat{R}_W + \hat{V}_W.$$

As before, equilibrium conditions in the employment and entrepreneurial markets exist only when the equalities in (3) hold.

## II. Entrepreneurs and Comparative Economic Welfare

Government programs to improve the standard of living for minority communities are typically judged on their impact on a variety of policy targets. The major target used in evaluating comparative welfare is the black/white income ratio (*BWIR*) defined as

$$(4) \qquad BWIR = \bar{Y}_B / \bar{Y}_W$$

where $\bar{Y}_B$ and $\bar{Y}_W$ are population mean incomes for the two groups.[5] Studies have shown that the *BWIR* is less than one and about .60.[6]

Concern on the part of policy makers with increasing the *BWIR* has led to a variety of programs including programs to promote minority entrepreneurship. Accepting the *BWIR* as the relevant measure of comparative economic welfare, we will now consider the entrepreneurial decision and changes in the *BWIR*. For ease in exposition we make the following definitions and assumptions.

[5] See James Gwartney, p. 872. This measure assumes that for any relevant geographical region in which there are $N$ and $M$ black and white individuals the ratio of means (or median income in the case of skewness) is, given that the standard deviation is identical for both income distributions, an adequate representation of relative economic well-being.

[6] Gwartney, pp. 873–74.

DEFINITIONS:

1. $\beta = W_B/W_W$
2. $\delta_i = W_i/(\hat{R}_i + \hat{V}_i), \qquad i = B, W$
3. $\Phi_t = Y_t^W/Y_t^B - 1$
4. $\theta_{ti} = R_{ti}/Y_{ti}, \qquad i = B, W$
5. $\alpha_t = (1 + r_t)/(1 + \rho_t)$

ASSUMPTIONS:

1. Consider a black and white worker of equal training and abilities. Because of labor market discrimination, $\beta < 1$ and $\Phi > 0$ for all t.

2. (a) Capital market restrictions are such that shifts from employment to entrepreneurship are allowed only if the prospective entrepreneur qualifies with some minimum value of $\delta(0 < \delta \leq 1)$, and (b) workers shift from employment to entrepreneurship at the maximum business values attainable, subject to capital market restrictions on $\delta$.

3. $\theta_t$ and $\alpha_t$ are constant for all t and their values will be expressed as $\theta$ and $\alpha$ respectively.

4. Homogeneous expectations and complete capital markets prevail.

*Proposition 1.* Consider a black worker and white worker of equal business training and ability. Because of discrimination in the labor market, the black worker can shift into entrepreneurship only at a smaller business value than that which is available to his white counterpart.

*Proof.* Define $f$ as the minimum value of $\delta$ required in the capital markets. Then $\hat{R}_W^* + \hat{V}_W^* = W_W/f$ and $\hat{R}_B^* + \hat{V}_B^* = W_B/f$ are the business values of each worker after shifting into entrepreneurship. Since $W_W > W_B$, then $\hat{R}_W^* + \hat{V}_W^* > \hat{R}_B^* + \hat{V}_B^*$, which completes the proof. The significance of proposition 1 is that even in the absence of ethnic discrimination in the capital markets a black cannot attain, on average, business value parity with a white of the same ability. An important reason is that financing institutions, such as banks, Small Business Investment Companies, require a minimum amount of equity capital to support a given size business. Earlier black capitalism programs *lowered* required equity values which resulted in adverse effects on the financing institutions as well as high default and business failures among minority entrepreneurs.[7]

*Proposition 2.* As black workers shift into entrepreneurship, the *BWIR* will increase, decrease, or remain the same depending upon the expected value of the firm, how this value is distributed across time, the magnitude of investment equity, and the relative size of $r_t$ and $\rho_t$.

*Proof.* The relationship between entrepreneurial income and employment income after a shift from employment to entrepreneurship can be expressed as

$$(5) \qquad \hat{R}_B + \hat{V}_B = \hat{Y}_B + W_B + \hat{C}$$

for the black worker, where $\hat{C}$ is some nonnegative present value. Dividing both sides by $\hat{Y}_B$ and simplifying with definitions 4 and 5 yields

$$(6) \qquad \sum_{t=0}^{T} \theta\alpha^t = 1 + \frac{W_B - \hat{V}_B + \hat{C}}{\hat{Y}_B}$$

It follows that

$$(7) \qquad \theta = \frac{1}{\sum_{t=0}^{T} \alpha^t} + \frac{W_B - \hat{V}_B + \hat{C}}{\hat{Y}_B \sum_{t=0}^{T} \alpha^t}$$

If $\theta = 1$, the *BWIR* remains the same; if $\theta > 1$, the *BWIR* increases, and if $\theta < 1$, the *BWIR* falls, which completes the proof.

The practical interpretation of proposition 2 is that the *BWIR* may *decrease* as blacks move from labor to entrepreneurial activities, even when there is a large in-

[7] See Bradford and Timothy Bates for an examination of this issue. A discussion of the magnitude of individual welfare losses from a Minority Enterprise Small Business Investment Company portfolio is found in Osborne.

crease in the present value of income ($\hat{C}$) when this shift is made. Three relationships are evident in the proof of proposition 2. When a shift is made by a black worker into the entrepreneurial position: (1) the higher the discount rate on entrepreneurial activity relative to the employment discount rate, the higher the positive change in the *BWIR*, (2) the larger the value of initial entrepreneurial wealth the higher the change in the *BWIR*, and (3) the higher the terminal value of the firm the lower the change in the *BWIR*, *ceteris paribus*. This final relationship is important in that a direct goal of governmental minority business programs is to encourage black business firms of higher present value ($\hat{V}_B$). As more emphasis is placed on viable businesses, withdrawals ($\hat{R}_B$) decrease, more funds are reinvested in the firm, and the *BWIR* falls.

Other relationships can be derived from the basic propositions. For example, if we assume "reverse discrimination" in some employment market (e.g., token blacks) such that $\hat{V}_B + W_B > \hat{V}_W + W_W$, then results opposite to proposition 1 hold. Such blacks can attain larger business values than whites of equal abilities and training.

Policy makers are no doubt concerned with the employment potential of business activity more than with the individual effects from one worker turned entrepreneur demonstrated in this paper. That extension to this model would recognize the possible multiplier effects of encouraging black entrepreneurs to hire marginal or unemployed black factors. As others have shown, this is not likely to occur if development is skill intensive.[8] Nevertheless, these are important issues which can also be developed beyond the basic framework briefly considered here.

[8] See Bates and Osborne on this point.

REFERENCES

T. Bates, "The Employment Potential of Inner City Black Enterprise," *Rev. Black Polit. Econ.*, Summer 1974, *4*.

W. Bradford and T. Bates, "Loan Default Amongst Black Entrepreneurs Forming Central City Businesses," *J. Fin.*, forthcoming.

G. S. Becker, *Human Capital*, New York 1964.

L. R. Christensen, "Entrepreneurial Income: How Does It Measure Up?" *Amer. Econ. Rev.*, Sept. 1971, *61*, 575–85.

J. Gwartney, "Changes in the Non-White/White Income Ratio 1939–67," *Amer. Econ. Rev.*, Dec. 1970, *60*, 872–83.

W. L. Hansen, *Education, Income and Human Capital*, New York 1970.

D. G. Johnson, "The Functional Distribution of Income in the United States, 1950–1952," *Rev. Econ. Statist.*, May 1954, *36*, 175–82.

S. Lebergott, "Factor Shares in the Long Term: Some Theoretical and Statistical Aspects," *The Behavior of Income Shares: Selected Theoretical and Empirical Issues*, National Bureau of Economic Research, *Studies in Income and Wealth, 27*, Princeton 1964.

G. M. Meier, *Leading Issues in Economic Development*, New York 1970, 5–9.

A. Osborne, "The Welfare Effects of Black Capitalists on the Black Community," *Rev. Black Polit. Econ.*, forthcoming, Summer 1976, *6*.

# Influences on Investment by Blacks in Higher Education

By WALTER W. MCMAHON*

There has been a significant improvement in the job markets for college educated blacks in the last 10 years, greater than the gains experienced by those who have not gone to college. National managerial job markets have opened up to blacks for the first time as evidenced by the major corporations recruiting blacks at the black colleges and by the increased number of governmental units and graduate schools seeking black graduates.

The additional years of college education are increasing black earnings but, even more important, are narrowing the earning gap between blacks and white. Median earnings of all blacks increased from 70 percent of median white earnings in 1967 to 81 percent of white earnings by 1974. The median earnings of black college graduates increased even farther to 94.7 percent of the earnings of white college graduates. (See T. F. Bradshaw 1975, p. 28.) This advantage of college training continued through the recession since unemployment rates in 1975 were 39–43 percent among those younger blacks with 8 years of education or some high school but only 3.3 percent among those blacks who finished college and had been in the labor force for two years. (See U.S. Bureau of Labor Statistics 1975, p. 5.)

This paper will present for the first time *ex ante* rates of return based on the earnings expected by individual black college students. Blacks and females are found to have unusually high expected rates of return, offering one reason for the increase in enrollments by blacks and females from 1970 through the 1975 recession. But since black and especially black male enrollments still remain disproportionately low, the access by black males to the newly open national managerial job markets remains limited by the influences that limit black enrollment in *U.S.* colleges. The second objective of this paper, therefore, will be to attempt to isolate the other major influences on the total amount of investment in college planned by blacks, and especially by black males. These influences, when compared to the investment behavior of white students and their families, can be useful in equalizing college enrollment persistence, and thereby access to the new job market opportunities that now await college educated blacks.

## I. Influences on Investment

The data consist of a nationwide sample of 7,019 prospective students and their families drawn in two waves in 1972 and 1973 in such a way that approximately half could have finished bachelor degrees in June 1975 and half in June 1976.[1] Blacks were deliberately oversampled to permit separate analysis of their behavior. The sample has been reweighted to conform to the distribution of all students with respect to sex, type of institution, and the

* Professor of Economics, University of Illinois at Urbana-Champaign. The author wishes to acknowledge research assistance from the Spencer Foundation and the National Institute of Education, and assistance with the collection of microeconomic data from the American College Testing Program.

[1] Questionnaires and further details of the survey are published in McMahon, pp. 167–79 and p. 23.

proportion receiving financial aid, as given by a census of all students.

### *A. Ex Ante Private Rates of Return*

The *ex ante* private rate of return was calculated for each student in the sample based on the earnings each student said he expected to receive at graduation and to earn 25 years later. Age-earnings curves of the shapes given by Hanoch then were fitted to the two points for each student before earnings at each age based on only a high school education were deducted. The resulting stream of net expected additions to earnings was set equal to the family's actual net tuition, foregone earnings, and other investment costs, and the internal rate of return was calculated.

It was found that blacks expect very high ex ante rates of return, ranging from 20 to 40 percent, and that the *ex ante* rates of return expected by white females were the next highest (see Table 1). Consistent with the hypothesis that investment decisions respond positively to high rates of return, enrollment by blacks grew the most rapidly nationwide in this period with black enrollments increasing from 7 percent of all college students in 1970 to 9 percent in 1974. (See U.S. Bureau of the Census 1975, No. 279, p. 10.) Relatively few black males were enrolled to begin with, and since black male enrollment has grown to only 8.6 percent of all college age males, with blacks constituting 12 percent of the college age population (U.S. Bureau of the Census, p. 15), black males remain significantly underrepresented.

The continuing underrepresentation of black males is also evidenced by the fact that only 26 percent of all black male high school seniors were planning to attend college last fall, in contrast to the much higher 41 percent of all white male high school seniors that were planning to attend. (See U.S. Bureau of the Census 1975, No. 284, p. 10.) So although high expected returns have induced increased enrollment of black males, their enrollment has started from modest levels, and other factors limit the amount of investment in education they plan to make. These in turn become the main remaining barriers to full access to national managerial job markets.

TABLE 1—EXPECTED RATES OF RETURN AND ENROLLMENT

| Those Enrolled in *BA, BS* Programs | *Ex-Ante* Private Monetary Rate of Return | | Percent Increase in Enrollment, 1970–74 |
|---|---|---|---|
| | 1975 Graduates | 1976 Graduates | |
| Black male | 37.6% | 19.2% | 66.8% |
| Black female | 31.7 | 40.5 | 45.7 |
| White male | 21.5 | 11.0 | 7.4 |
| White female | 24.2 | 32.6 | 26.7 |

### *B. Factors Limiting Black's Investment Plans*

To isolate some of the other influences on the investment decisions made by black students and their families, a two equation model composed of an investment-demand (or marginal efficiency of investment) function and a supply-of-resources function was estimated using this microeconomic survey data by three stage least squares simultaneous equation methods. The model was estimated separately for black females, black males, white females, white males, and for all families in the lower income quartile (see Table 2).

The first result found is that although high expected rates of return ($r$) have encouraged black enrollments, the marginal efficiency of investment schedule does slope downward for black and for other population segments as expected, indicating that longer and more costly planned degree programs yield lower rates of return. (The coefficient of $r$ is negative and significant in Equations (1), (3),

TABLE 2—INFLUENCES ON THE AMOUNT OF INVESTMENT PLANNED IN HIGHER EDUCATION

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | *Black Females* (June 1976 B.A.s) | | | | | |
| *Total family investment, black females:* | | | | | | | | | |
| (1) Investment demand ($MEI$): | $I_F=$ | $\underline{-14.7\,r}$ | $\underline{-9.14\,\mu}$ | $\underline{+.25\,A}$ | $\underline{-.16\,S_p}$ | $+2.41\,\alpha_c$ | $\underline{+19.45}$ | | |
| | [a] | $(-3.27)$ | $(-3.84)$ | $(2.41)$ | $(-2.13)$ | $(.65)$ | $(3.51)$ | | |
| (2) Supply-of-resources: | $I_F=$ | $-.62\,r$ | $\underline{+.12\,Y}$ | $\underline{+2.27\,S}$ | $\underline{+.56\,L}$ | $\underline{-.47\,W}$ | $\underline{-.44\,N}$ | $\underline{-.47\,Y_o}$ | $\underline{+11.49}$ |
| | | $(-.99)$ | $(3.08)$ | $(6.04)$ | $(10.71)$ | $(-9.66)$ | $(-7.19)$ | $(-3.55)$ | $24.73$ |
| | | | | *Black Males* (June 1976 B.A.s) | | | | | |
| *Total family investment, black males:* | | | | | | | | | |
| (3) Investment demand ($MEI$): | $I_F=$ | $\underline{-3.06\,r}$ | $-.18\,\mu$ | $+.12\,A$ | $\underline{+.63\,S_p}$ | $+7.61\,\alpha_c$ | $-18.3$ | | |
| | [a] | $(-2.37)$ | $(-.05)$ | $(1.01)$ | $(4.48)$ | $(1.39)$ | $(-1.76)$ | | |
| (4) Supply-of-resources: | $I_F=$ | $\underline{-3.48\,r}$ | $-.004\,Y$ | $\underline{+3.18\,S}$ | $\underline{+.63\,L}$ | $+.003\,W$ | $\underline{+.33\,N}$ | $\underline{+.73\,Y_o}$ | $\underline{+11.53}$ |
| | | $(-4.31)$ | $(-1.15)$ | $(5.69)$ | $(10.65)$ | $(.30)$ | $(3.29)$ | $(5.71)$ | $(18.92)$ |
| | | | | *White Males* (June 1976 B.A.s) | | | | | |
| *Total family investment, white males:* | | | | | | | | | |
| (5) Investment demand ($MEI$): | $I_F=$ | $\underline{-36.\,r}$ | $\underline{-6.2\,\mu}$ | $\underline{+.29\,A}$ | $\underline{+.16\,S_p}$ | $\underline{+8.37\,\alpha_c}$ | $\underline{+4.62}$ | | |
| | [a] | $(-7.9)$ | $(-2.5)$ | $(2.61)$ | $(2.01)$ | $(2.05)$ | $(.85)$ | | |
| (6) Supply-of-resources: | $I_F=$ | $+.43\,r$ | $\underline{+.07\,Y}$ | $\underline{+.90\,S}$ | $\underline{+.79\,L}$ | $\underline{-.75\,W}$ | $\underline{+.72\,N}$ | $\underline{+.265\,Y_o}$ | $\underline{+13.7}$ |
| | | $(1.21)$ | $(4.68)$ | $(4.55)$ | $(30.8)$ | $(-48.6)$ | $(16.7)$ | $(6.45)$ | $(65.1)$ |
| | | | | *Black Females* (June 1976 B.A.s) | | | | | |
| *Parents contribution only, black females:* | | | | | | | | | |
| (7) Demand ($MEI$): | $I_F=$ | $.67\,r$ | $+1.34\,\mu$ | $+.07\,A$ | $-.04\,S_p$ | $-37.5\alpha_c$ | $+4.76$ | | |
| | [a] | $(.27)$ | $(1.02)$ | $(1.32)$ | $(-.84)$ | $(-1.85)$ | $(1.56)$ | | |
| (8) Supply-of-resources: | $I_F=$ | $-.26\,r$ | $\underline{+.13\,Y}$ | $+.17\,S$ | $\underline{-.18\,L}$ | $\underline{-.08\,W}$ | $\underline{-.23\,N}$ | $\underline{-.42\,Y_o}$ | $\underline{+3.88}$ |
| | | $(-.79)$ | $(6.22)$ | $(.88)$ | $(-6.59)$ | $(-3.14)$ | $(-7.37)$ | $(6.20)$ | $(16.05)$ |
| | | | | *Lowest Income Quartile* (All Races, June 1975 B.A.s) | | | | | |
| *Total family investment, blacks and whites, males and females:* | | | | | | | | | |
| (9) Demand ($MEI$): | $I_F=$ | $\underline{-.32\,r}$ | $-.08\,\mu$ | $\underline{-6.36\,\text{Race}}$ | $\underline{-3.43\,\text{Sex}}$ | $+.23\,A$ | $-.13\,S_p$ | $\underline{+28}$ | |
| | [a] | $(-6.38)$ | $(-.03)$ | $(-3.08)$ | $(-2.32)$ | $(1.48)$ | $(-1.03)$ | $(5.72)$ | |
| (10) Supply-of-resources: | $I_F=$ | $\underline{-.015\,r}$ | $\underline{-.27\,Y}$ | $\underline{+.61\,L}$ | $\underline{-.34\,S'}$ | $\underline{-5.68\,W}$ | $-.07\,N$ | $\underline{+11}$ | |
| | | $(-4.01)$ | $(-3.46)$ | $(10.93)$ | $(-2.84)$ | $(-22.27)$ | $(-1.02)$ | $(25.63)$ | |

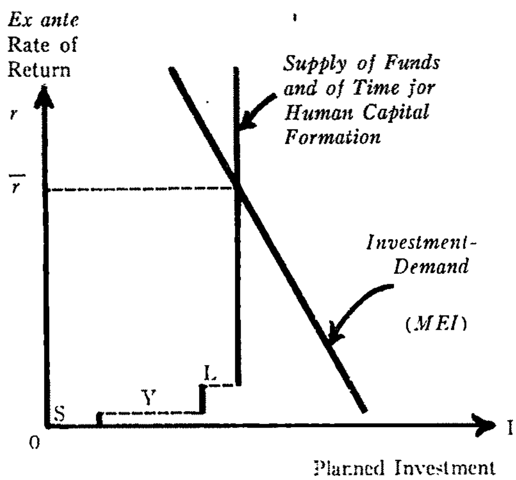
*Note*—investment demand ($MEI$) and supply-of-funds coefficients were estimated simultaneously for each population segment by three stage least squares. Respondents could have finished $BA$'s in June 1975 or June 1976. Coefficients significant at the .05 level are underlined, and $t$-statistics are in parentheses.

[a] For three stage least squares estimation, $\hat{I}$ is moved to the left in each supply equation.

$\alpha_c$ = A taste parameter reflecting the importance of college as an aid in making a choice of a career. Other variables are defined in the text.

(5), and (9), which is reflected in the negative slope of the investment-demand function in Figure 1.)

Second, uncertainty about future earnings ($\mu$) which might be expected to curtail student investment plans does limit the amount of investment planned by white males (5) and black females (1), but less significantly by black males (3). But third, higher ability ($A$) as measured by the American College Testing Service ($ACT$) achievement test scores clearly is a significant determinant encouraging most whites and black females to make more ambitious plans as predicted by economic theory. Ability is a less significant factor for black males, perhaps because black families more commonly expect males to go to work after high school. For black males, the number of years of schooling of the parents ($S_p$) however is is a uniquely significant positive determinant (in 3) of whether black males plan to go further in college.




I = Tuition and fees, books, and foregone earnings through planned degrees.

FIGURE 1. INVESTMENT DECISIONS MADE BY BLACK FAMILIES

It is on the supply-of-resources side that most of the highly significant effects on the size of the investment planned by black families are to be found. The most significant positive effects are from scholarships ($S$) and student loans ($L$) supporting larger student investment plans in (2), (4), and (6). The second largest positive effect is from family disposable income ($Y$), especially for females in (2) and (8). This significance of family income and related alternative sources of financial support is consistent with other studies using both microeconomic and time series data that also normally find income sources to be fundamental (e.g., McMahon 1974). A third significant factor on the supply side is part-time work ($W$) which by reducing the amount of time blacks (and others) invest in study is associated with persistence but with lower total planned investment.

In all the regressions tried, loans reduce the *parental* contribution in both black (8) and white families, but they increase the amount of investment planned by the student so that the net effect on total family investment is positive. See (2) and (4). A larger number of siblings in the family ($N$) clearly curtails the parents' contribution (8) which limits the resources available and significantly curtails the amount of parental support for both black and white females, as might be expected.

In summary, the most significant influences on college investment plans made by black males on the demand side are the newly held expectations of high rates of return ($r$) and tastes acquired from parents who have been to college ($S_p$). Improvements in the quality of primary and secondary education for blacks may also do much to improve ability test scores ($A$). But access to the occupations where the rates of return are highest such as business management, medicine, pharmacy, engineering, and law requires years in college. And investment in college by black males especially continues to be restricted by their primary and secondary school backgrounds ($A$) and by more limited sources of funds from parents—$Y$ in (4) and (10)—who themselves have not been to college.

With the improvements in the last decade in the opportunities available to college educated blacks, these influences become very significant factors limiting access by black males to the national managerial job markets.

## REFERENCES


T. F. Bradshaw and J. F. Stinson, "Trends in Weekly Earnings," *Monthly Labor Rev.*, Aug. 1975, *98*, 22–32.

W. W. McMahon, *Investment in Higher Education*, Lexington 1974.

U.S. Bureau of the Census, *Current Population Reports*, Series P–20, No. 279 (March) and 284 (Sept.), Washington 1975.

U.S. Bureau of Labor Statistics, *Summary, Special Labor Force Reports*, "Educational Attainment of Workers," Washington 1975.

# Some Alternatives for Reducing the Black-White Unemployment Rate Differential

*By* Karl D. Gregory*

Economic development involves promoting higher standards of living in part through generating saving for adding to the capital stock and productive potential of a community. Saving is a function of income. Most income is obtained from wages and salaries. It is therefore quite appropriate that this American Economic Association-National Economic Association session on "Trend and Analysis in Black Economic Development" include a paper on recent changes in the work status of blacks. Such a focus has been made more timely by the recent recession during which unemployment rates peaked at 8.9 percent in 1975–II, and for blacks, at 14.3 percent. Making matters worse, this was accompanied by a rate of inflation in the recession, the deepest of all since the Great Depression, which rose to double digit levels in 1974 and was over 7 percent in 1975.

Measurement of the unequal work status among blacks and whites in the United States involves, among other factors, differences in labor force participation rates, unemployment rates, discouraged worker rates, underemployment rates, working conditions, and levels of wages, fringe benefits and other compensation. This paper discusses primarily only one attribute of work status, unemployment, despite the incompleteness of this concept. (Other research in which I am currently engaged, along with associates in the Congressional Budget Office, will analyze the other attributes.) The paper summarizes the recent behavior of the black-white unemployment differential, reviews the likely impact on the differential of two recovery paths, one for moderate recovery and another for rapid recovery by 1978–IV, and briefly suggests some alternative selective programs for narrowing the black-white differential in a short time period. I shall try to share with you some of the tentative and preliminary conclusions on these matters produced by ongoing research at the Congressional Budget Office.

## I. Unemployment and the Black-White Differential, Recent Experience

The extent to which the black unemployment rate exceeds that for whites (hereafter referred to as the gap) varies greatly over the business cycle. For the last 20 years this gap has ranged from 2.9 percentage points in 1969 to 6.8 percentage points in 1958. Not surprisingly, it is lowest at the end of an extended economic ex-

* Visiting Scholar, Congressional Budget Office (*CBO*), and Professor of Economics and Management, Oakland University. The author is indebted to colleagues at *CBO* with whom he has been conducting studies on reducing differences in the status of blacks and whites in the United States. He alone is responsible for any errors. Charts depicting relationships discussed in the text are available from the author. Two publications by *CBO* discuss the subject more extensively: "The Impact of Economic Recovery on Unemployed Nonwhite and White Americans: A Preliminary Assessment," (mimeo) and *Temporary Measures to Stimulate Employment: An Evaluation of Alternatives,* 1975.

TABLE 1—UNEMPLOYMENT RATES BY RACE, SEX AND AGE DURING THE CURRENT RECESSION

| Demographic Group | Unemployment Rate[b] (in percent) 1973–IV | Unemployment Rate[b] (in percent) 1975–III | Jobless Rate[c] (in percent) 1975–III |
|---|---|---|---|
| All | 4.7 | 8.4 | 9.5 |
| Adult males | 3.1 | 6.9 | |
| White | 2.8 | 6.4 | NC[d] |
| Nonwhite[a] | 5.4 | 11.5 | |
| Adult females | 4.7 | 7.7 | |
| White | 4.2 | 7.0 | NC |
| Nonwhite[a] | 8.1 | 11.8 | |
| Teen-agers (16–19) | 14.3 | 19.8 | NC |
| White | 12.8 | 18.0 | |
| Nonwhite[a] | 28.3 | 36.0 | |
| All whites | 4.3 | 7.7 | 8.7 |
| All nonwhites | 8.6 | 13.8 | 16.4 |

[a] Negro and other races of whom 89 percent are Negro. See *Employment and Earnings: October 1975*, U.S. Department of Labor, Bureau of Labor Statistics, p. 161.

[b] Computed from seasonally adjusted Bureau of Labor Statistics *News* data, *op. cit.*

[c] Defined for each demographic grouping as the unemployed plus discouraged workers divided by the labor force plus discouraged workers.

[d] Not computed.

pansion, when the rate for adult white males tends to be explained completely by frictional unemployment, and highest at or near the trough of the business cycle. In the current cycle, the overall gap in unemployment rates has widened from 4.3 percentage points in 1973–IV to 6.1 in 1975–III. The gap of 6.1 in 1975–III would widen to 7.7 percentage points with the inclusion of discouraged workers, and would be even larger with further adjustments for including the underemployed. With the inclusion of the discouraged worker (Table 1) but not the underemployed, the jobless rates in 1975–III were 8.7 percent for whites, as compared to 16.4 percent for blacks.

The pattern of the racial gap over the last two decades has also varied by sex and age. For adult males (aged 20 and over), while unemployment rates for whites have varied between 6.6 and 1.8 percent, those for blacks have ranged from 13.9 to 3.6 percent. The resulting gap for adult males has fluctuated between 7.8 and 1.6 percentage points.

The greatest reduction in the gap for males resulted at the end of the sustained expansion throughout most of the 1960's. Virtually all of these gains appear to have been wiped out in the early 1970's, for the gap for males has returned to the levels of the mid–1950's. Indeed, the recent decline in labor force participation rates for nonwhites suggests that the increased gap in the unemployment rate understates the extent to which the gains of the 1960's have been erased, for many potential labor force members have given up seeking work, believing jobs to be unavailable. Black males are disproportionately represented among these discouraged workers who are not counted as being unemployed and therefore are not reflected in the gap.

For adult females, while the unemployment rate for whites has varied from 8.0 to 3.2 percent, that for blacks has ranged from 11.8 to 5.4 percent. The unemployment gap had been lower for females than for males prior to the mid–1960's but became larger than for males in the late 1960's as men gained more than women from the prolonged expansion in that period. By mid–1975, the gains in the reduced gap for females were also extinguished.

For teenagers, the gap in unemployment rates has been continuously worsening over the two decades. While the white teenage unemployment rate has fluctuated between 18.3 and 9.1 percent, the black rate has varied between 39.8 and 13.5 percent. The gap has grown from 4 percentage points in the mid–1950's, to 14 in the mid–1960's, to 20 in the mid–1970's. The economic ebullience of the 1960's merely tempered the persistent trend up-

ward in the black-white teenage unemployment gap.

## II. The Pace of Recovery and the Unemployment Gap

A central question then is what can be done to reduce unemployment rates generally, as well as the gap between the rates for blacks and whites. It is therefore of interest to ascertain the contribution that might be made by a more rapid recovery than the modest ones often contemplated in current popular forecasts.

Since the recent recession has been unusually severe, a return from current rates of unemployment of above 8 percent to a low level by 1978–IV, say to 4.5 percent, would require very high average annual rates of growth in the real *GNP* approximating 7.5 percent. This in turn would be consequent to highly expansionary fiscal and/or monetary policy or to a vigorous upswing in demands from the private sector, for example, a decline in the saving rate, or a boom in capital spending or in net exports. There is no evidence of such strong sources of forthcoming private demand. A recovery averaging a real growth rate of 5.5 percent annually is more consistent with past experience over a three year period. Such a growth rate would leave the unemployment rate at between 6 and 7 percent by the end of 1978. There are of course other possible growth paths for the economy.

The trade-off with inflation from more rapid rates of recovery is difficult to project. Because of the persistence of inflation, it is quite probable that even the more moderate recovery would still be accompanied by annual increases in the consumer price index of about 6 percent by 1978. Accelerating the average annual growth rate from 5.5 percent to 7.5 percent until 1978 through stimulative fiscal policy with less than fully accommodating monetary policy could increase the rate of inflation by another percentage point. It can be argued that as long as there is excess capacity and unemployment remains about 4.5 percent, added inflationary pressures will be moderate. On the other hand, as our report states, "the psychological consequences of adopting a new, highly expansionary fiscal strategy are uncertain, and one cannot rule out the possibility of a new round of price increases that could be set off by inflationary expectations." The depth of the current recession and the concurrence of high unemployment together with inflation combine to increase the hazards of applying models based on past experience to the current recovery.

TABLE 2—PROJECTED RANGE OF SELECTED UNEMPLOYMENT RATES FOURTH QUARTER, 1978

| | Alternative Macroeconomic Policies | |
|---|---|---|
| | Faster Recovery Strategy 1978–IV | Moderate Recovery Strategy 1978–IV |
| Unemployment rate (in percent) | | |
| Total | 4.5 | 6.6 |
| Nonwhites[a] | 8.0–8.8 | 11.4–12.2 |
| Whites | 3.9–4.1 | 5.8–6.0 |
| Difference between nonwhite and white unemployment rates (in percentage points) | 3.9–4.9 | 5.4–6.4 |

*Source:* The unemployment rates and differentials were estimated by *CBO*, assuming a continuation of recent historical patterns. The levels of the nonwhite and white unemployment rates relative to the national unemployment rates were reviewed and particularly for recent periods when the national rate approached those projected for both recovery paths. The estimates were also compared to the mid-points produced in two models using microsimulation modeling of the labor market at the unemployment rates of 4 and 5 percent at 1978–IV. The microsimulations were made available to *CBO* by The Urban Institute.

[a] Negro and other races, 89 percent of whom are Negro. See *Employment and Earnings: October 1975*, U.S. Department of Labor, Bureau of Labor Statistics, p. 161.

Shown in Table 2 are the unemployment rates by race and the racial gap consistent with the two recovery paths. The moderate recovery path leads in 1978–IV to a national unemployment rate of about 6.6 percent. The associated range of unemployment rates for whites has a midpoint at 5.9 percent in contrast to 11.8 percent for blacks, leaving a racial gap of between 5.4 and 6.4 percentage points in 1978–IV.

A more rapid recovery could be brought about by cuts in the personal income tax, increased government purchases of goods and services, or by a combination of the two at a cost of a greater federal deficit and, as we have seen, added inflation. It would produce a 4.5 percent unemployment rate in 1978–IV, according to *CBO* simulations. In contrast to the moderate recovery strategy, this strategy would reduce the overall unemployment rate in 1978–IV by 2.1 percentage points. The midpoint of the range of associated unemployment rates for whites would be lowered by 1.9 percentage points, and for blacks, by 3.4 percentage points. More rapid recovery would therefore narrow the gap by about 1.5 percentage points, or by about one-fourth, a not inconsiderable magnitude. Even so, a gap ranging from 3.9 to 4.9 percentage points would still remain between the black and white unemployment rates in 1978–IV.

## III. Conclusion

Over the last 20 years, the gap between the unemployment rates of blacks and whites has ranged from 2.9 to 6.8 percentage points and was close to the high end of this range in 1975–III when the gap was 6.1 percentage points. The gap varies by sex and age. The substantial gains made in the 1960s in reducing the racial gap for adult males and for adult females had been virtually wiped out by mid–1975. For teenagers, the gap has become progressively wider over the last twenty years. Adjusting unemployment rates to include the effects of discouragement would increase the overall racial gap in 1975–III from 6.1 to 7.7 percentage points. No adjustments were made to adjust the data for underemployment.

A major factor affecting the level of unemployment and perhaps the gap is the pace of the recovery. Two growth paths over the next 36 months were reviewed. The real *GNP* expands at an average annual rate of 5.5 percent in the moderate growth model and by 7.5 percent in a faster growth model. The moderate growth strategy reduces the overall unemployment rate from 8.3 percent in November 1975 to 6.6 percent by 1978–IV, but it would either not close the racial gap at all or not by much. The faster growth strategy reduces the unemployment rate to 4.5 percent by 1978–IV and narrows the racial gap by one-fourth, at a cost of some added inflation and a higher deficit. Even with the faster growth strategy, a gap between the black and white unemployment rates of between 3.9 and 4.9 percentage points remains unclosed.

Ongoing research is focusing upon alternative selective fiscal policies for further narrowing the gap. Alternatives such as countercyclical revenue sharing, public works, public service employment and other public programs that are targetable will be analyzed for their impact on the gap in the short run and their costs. Other research at the Congressional Budget Office analyzes what is known about the effectiveness of a wider range of programs—including job creation strategies, human capital development and regulatory efforts aimed at combatting discrimination—that would impact on the racial gap over a longer time period.

# Capital Problems in Minority Business Development: A Critical Analysis

*By* KENNETH E. KNIGHT AND TERRY DORSEY*

The authors have examined the federal government's strategy of economic development for minority groups through the development of minority enterprise as a function of capital availability. (See Knight and Dorsey for complete analysis.) The demand for capital is defined as a function of asset parity with nonminority business to be achieved over a twenty year period. Capital availability is estimated through examination of three sets of institutions which provide capital to minority business: (1) minority and nonminority banks; (2) the availability of equity capital through the Minority Enterprise Small Business Investment Company (*MESBIC*) program; and (3) the Small Business Administration (*SBA*).

Based upon an estimate of the annual capital requirements needed to achieve asset parity over time, the magnitude of available capital is assessed for adequacy. In aggregate, the three sets of institutions appear to provide less than 10 percent of the yearly amount required to reach asset parity. This capital shortfall is viewed as the result of a more generalized capital shortage in the United States today.

Between 1960 and 1973, the *U.S.* average level of total output devoted to fixed investment was lower than in any of the other 11 most industrialized countries, including Italy and Great Britain (U.S. Treasury Department). Estimates of capital shortage expected over the next ten years range from approximately $200 billion by James S. Duesenberry and Bosworth (*Business Week*) to $646 billion by a New York Stock Exchange study. While it is not possible to estimate the impact of this shortage for minority enterprise, it is appropriate to conclude that the shortage can do little except make the lack of available capital for minority enterprise even worse. The current programs, while significantly larger than efforts in the past, fall far short of what is required to achieve asset security within 20 years.

## I. The Demand for Capital by Minority Business

Estimating the demand for capital by minority business from the goal announced by Secretary Stans in 1970—double the number of minority firms between 1970 and 1990—is inadequate as an operational goal. The inadequacy of this goal stems from its failure to enumerate the types and sizes of firms that would most benefit minority groups. R. Yancy describes the results of an unfocused program to assist minority enterprise:

> The types of industries which blacks have entered under federal assistance programs are no different from the industries they enter without federal assistance. These industries . . . do not offer optimum opportunities for economic growth and success. Few of the industries where blacks are concentrated offer excellent opportunities for large scale activity, for capital accumulation or for growth . . .

* The University of Texas at Austin.

TABLE 1

| Firm Size (Yearly Receipts) | Receipts Per Employee | |
|---|---|---|
| \$0–\$49,999 | \$12,269 | |
| \$50,000–\$99,999 | \$21,198 | |
| \$100,000–\$199,999 | \$25,191 | |
| \$200,000 or more | \$36,570 | |
| | Total *U.S.* Firms | Minority Firms |
| Number of all firms | 7,489,000 | 322,000 |
| Gross sales | \$1,498 bil. | \$10.6 bil. |
| Average sales | \$200,027 | \$32,919 |
| Total nonfinancial (corporate) assets | \$859 bil. | \$2.6 bil. |

> The ideal federal strategy, then, should be to encourage blacks to enter larger scale businesses which employ greater numbers of employees and which offer potential for growth within the business community. [pp. 67–68]

Potential benefits accruing to the minority community from development of larger ventures rather than small or "mom and pop" operations is highlighted by data from the 1969 census of minority-owned businesses with paid employees (p. 152). (See Table 1.)

If one asserts that the objective of minority business development is parity of total minority business assets over a twenty year period, it is necessary to make a yearly capital investment of approximately \$7.3 billion. (Minorities constitute approximately 17 percent of the *U.S.* population; \$859 billion x 1/20 x 17% = \$7.3 billion.) The overall debt/equity ratio for all nonfinancial *U.S.* corporations in 1971 was 2.57 (*Statistical Abstract of the U.S.*, 1974); thus, it may be inferred that approximately \$2.04 billion should be invested or generated each year as equity and \$5.24 billion as debt.

## II. The Availability of Capital for Minority Enterprise

Next the supply of capital for minority business relative to the demand is evaluated.

### A. The Banking System

Examination of capital availability for minority business through commercial banks is conducted by analyzing the roles of minority and nonminority banks.

Andrew Brimmer succinctly describes characteristics of black banks: "The typical black bank is about one-third the size of the typical bank in the country at large, the former having average deposits of roughly \$10 million in 1969 compared with \$33 million for the latter . . ."

Other factors diminish black banks' ability to loan capital to business even further. High losses and inexperienced management have combined to produce (1) high operating costs and low efficiency, (2) one-quarter to one-third the percentage profit of the average bank, (3) loans of a smaller proportion of the bank's available funds, and (4) investments of a large proportion of total assets in government securities.

Despite the almost insignificant size of minority banking relative to the commercial banking system as a whole, its role through 1970 in financing minority businesses

> . . . had been disproportionately large: As recently as 1971, it has been reported, minority banks, with combined assets less than one tenth of one percent of those of all commercial banks, had extended \$60 million in loans to minority enterpreneurs —more than a third as much as the \$150 million granted by nonminority banks. [D. Fisher, p. 37]

However, in 1970 the banking industry announced a program to invest \$1 billion in the new financing of businesses planned or owned by members of minority groups by 1975 (*Burroughs Clearing House*). Statistics released by the American Bankers Association (*ABA*) in July

1973 revealed that 351 member banks had placed minority loans totaling $758 million since 1970. While lack of suitable data prohibits precise comparisons, it is useful to note that $758 million represents 2.93 times $258.8 million, the total assets of black banks in 1969 and 9.12 times $83.1 million, the total amounts of loans by black banks for commercial, industrial and real estate purposes in 1969. However, if $758 million were invested in only one year, it would represent 14 percent of the amount of debt capital needed in that year to achieve parity over a twenty year period.

### *B. Availability of Equity Capital: The MESBIC Program*

The availability of long-term debt and equity capital for minority business is examined by analyzing the *MESBIC* and, to a lesser extent, the Small Business Investment Corporation (*SBIC*) program. This approach is undertaken because minority groups own negligible financial assets to serve as the basis of equity capital.

By April 1972, *MESBIC*s had invested $7.6 million in 422 firms, averaging $18,009 per firm (Richard H. Klein); assuming 6 percent annual inflation, average investment size in 1975 dollars would be $21,449. In February 1975, there were 73 licensed *MESBIC*s whose combined private and government capital was $68.1 million.

Experience from the *SBIC* indicates that smaller investment firms are more prone to make debt rather than equity investments. Because of their small sizes, it is estimated that firms in the *MESBIC* program probably can furnish no more than $14.73 million in equity funds *if they were fully invested.* Using an estimated average investment size of $21,500, it is inferred that the current *MESBIC* program is likely to furnish equity capital to approximately 650 ventures.

The entire *MESBIC* program, consisting of $68.1 in capital equals approximately 3.5 percent of the yearly need for almost $2 billion in equity. If the *SBA* requirement of 15 percent equity is used, the entire *MESBIC* program can support $454 million in *SBA* loans or 6.25 percent of the needed yearly investment.

Samuel Doctors and S. Lockwood have proposed that entry by minority enterprise into growth industries, especially technology-intensive ones, would permit the most economic benefits to minority enterprise. Thus, it is useful to evaluate the availability of venture capital (*VC*) and to note that the small size of the *MESBIC* industry derives not so much from inadequate support by government as it does from the total amount of funds currently available as equity capital for new nonminority enterprise.

Stanley Rubel, editor of *Venture Capital,* estimates that the approximately 600 *VC* operations listed in his directory have access to about $400 million which can be invested as *VC;* he also estimates that these operations will invest approximately $150–250 million this year. If in 1975 the entire *VC* community made all its investments in minority enterprise, $250 million would still represent only 12.3 percent of the amount needed this year to achieve asset parity. Moreover, the total $400 million estimated to be available is only 19.6 percent of the $2.04 billion needed each year as equity.

Just as smaller loans made by *SBA* seem to have higher probabilities of failure, evidence also indicates that large *VC* investments have higher probabilities of success than smaller ones. A study of more than 350 such investments reveals that investments which resulted in very successful growth ventures averaged more than $800,000 each, while those which produced failures averaged $400,000 (Dor-

TABLE 2—MINORITY ENTERPRISE LOANS
(Millions of dollars)

| Year | Total *SBA* Involvement[a] | *SBA* Direct (37.5%) |
|---|---|---|
| 1968 | $ 54 | $ 20.25 |
| 1969 | 131 | 49.13 |
| 1970 | 176 | 66.0 |
| 1971 | 231 | 86.63 |
| 1972 | 309 | 115.88 |
| 1973 | 317 | 118.88 |
| 1974 | 238 | 89.25 |

[a] 1968–72 from Yancy, p. 72. 1973–74 from 1974 Annual Report of *SBA*.

sey). Because of statutory limitations on maximum investment size as a function of private capital invested, there are no *MESBIC*s which can make investments of even $400,000 or more, much less the larger investments.

Thus, the availability of equity capital through the *MESBIC* program is inadequate: (1) the size of the program is too small; (2) the size of individual *MESBIC*s is likewise too small to finance firms with high probabilities of success. However, the effectiveness of *MESBIC*s is limited not by the program itself as much as by the limited availability of equity capital for all new firms.

### *C. Small Business Administration*

Analysis by Yancy indicates that direct investment by the *SBA* in business is somewhat less than claimed because *SBA* includes participation loans and loan guarantees when it publishes data on loan investment. Using Yancy's data to adjust published *SBA* data, the authors are able to estimate the amount of capital actually loaned to minority business between 1968 and 1972. See Table 2.

The $238 million *SBA* claims to have invested in minority business in 1974 is 4.47 percent of the yearly amount required to achieve parity over time; the adjusted estimate of $89.25 is only 1.67 percent of the needed yearly amount.

## REFERENCES

T. **Bates**, "An Econometric Analysis of Lending to Black Businessmen," *Rev. Econ. Statist.*, Aug. 1973, 272–83.

A. **Brimmer**, "The Black Banks: An Assessment of Performance and Prospects," *J. Fin.*, May 1971, 379–402.

**Burroughs Clearing House**, "Banks Commit $1 Billion in Loans to Minority Groups," June 1970, 12–13.

**Business Week**, "Where Will the Growth Capital Come From," Dec. 14, 1974, 70–71.

S. **Doctors**, *Whatever Happened to Minority Economic Development?* Hinsdale 1974.

——— **and S. Lockwood**, "New Directions for Minority Enterprise," *Law and Contemporary Problems, 36*, 1971.

T. **Dorsey**, *Factors Responsible for Success in the Investment of Venture Capital*, unpublished doctoral dissertation, University of Texas at Austin 1975.

D. **Fisher**, "Minority Banks: A Progress Report," *Conference Board Record*, Dec. 1973, 37–41.

A. **Karana-Karan and E. Smith**, "A Constructive Look at MESBICs," *California Management Review*, Spring 1972, *14*, 82–87.

R. **Klein**, "A Perspective on the MESBIC Program," *Michigan State University Business Topics*, Autumn 1972, 45–51.

K. **Knight and T. Dorsey**, "Capital Problems in Minority Business Development," Working pap., 76–14, Bureau of Business Research, University of Texas at Austin.

**New York Stock Exchange, Inc.**, "The Capital Needs and Savings Potential of the U.S. Economy, Projections through 1985," New York, Sept. 1974.

S. **Rubel**, telephone interview conducted Nov. 11, 1975.

———, *Guide to Venture Capital Sources*, Chicago 1974.

**U.S. Bureau of the Census**, *Statistical Abstract of the United States*, 1974, 483.

**U.S. Dept. of Commerce**, *Minority-Owned Business: 1969*, Washington 1971.

**U.S. Treasury Dept.**, "Department of Treasury News," Washington, Apr. 1, 1975, 103.

R. **Yancy**, *Federal Government Policy and Black Business Enterprise*, Cambridge, Mass. 1974.

# INTERNATIONAL CAPITAL MARKETS AND FOREIGN EXCHANGE: NEW DIRECTIONS OF RESEARCH

# Portfolio Choice and the Demand for Forward Exchange

*By* MICHAEL ADLER AND BERNARD DUMAS*

The development of the so-called "modern theory" (*MT*) of the forward exchange market summarized in Herbert G. Grubel and Hans R. Stoll (1968) purported to offer at least three contributions. First, it established that forward exchange market equilibrium could occur at a forward rate which did not correspond to interest rate parity (*IRP*). Second, it predicted usefully for forecasting purposes that the equilibrium forward rate would invariably fall in between the *IRP* forward rate and the expected future spot rate. And third it supplied a theoretical foundation for government policies of intervention in the forward market which were deemed—given certain assumptions regarding speculators' expectations as to the results of the intervention—to be potentially successful even under market demand conditions that were such as to produce *IRP*. As we shall show, the first claim survives but for reasons other than those proposed by the *MT*, while the second and third do not.

The main conceptual difficulties of the *MT* deserve brief mention. The urge to identify the equivalent of demand and supply sides in the foreign exchange market forced many authors, e.g., Sho C. Tsiang, Egon Sohmen and Grubel, to posit an artificial segmentation of forward market transactions as between pure arbitrage and pure speculation: arbitrageurs never speculate and vice versa. Because interest arbitrage involves equal and opposite spot and forward transactions, this putative segmentation led some to the misleading supposition of a direct quantitative link between the volumes of forward and spot market transactions and an incorrect specification of the response of spot, short-term capital flows to monetary policy.[1] The segmentation also forced the *MT* largely to exclude hedging, an important motive of transaction, for which the linkage necessarily breaks down (E. Ray Canterbery, Donald Schilling). But the exclusion of hedging is actually unnecessary. The link is generally indirect and will be established by the simultaneous equilibria in the international markets for stocks, bonds and currency.

Our main objective in Section I is to derive an individual's demand function for forward exchange under general uncertainty in a two-country model which includes international markets for stocks, bonds, and forward contracts and which allows for default risks. Peter B. Kenen, using a certainty model of an individual

* The authors are Professor and Assistant Professor, respectively, at the Graduate School of Business, Columbia University. Financial support was provided by the Rockefeller Foundation which, however, bears no responsibility for any opinion expressed herein.

[1] Jerome L. Stein provides an exception to this rule.

firm, was the first to argue that a single economic unit might trade, hedge, and speculate all at once. Similarly, our investors, who may trade to an unspecified degree, will transact simultaneously in the several financial markets.

Our analysis of Section II reveals that the demand for forward contracts cannot be decomposed linearly into additive component demands for the purposes of arbitrage and speculation. The usual *MT* rendition is therefore demonstrably wrong. The arbitrage motive operates multiplicatively: it amplifies the demands for other motives and need not involve interest arbitrage at all. Arbitrage will occur between *any* two linearly dependent sets of risky assets. Further, a demand for hedging purposes, excluded from the *MT*, arises here as a result of the availability of forward contracts as an extra risky asset into which the investor can diversify to reduce his total real portfolio risk. Perhaps surprisingly, hedging in our model has nothing to do with definitions of the exposure of assets or investment returns to exchange risk. An investor maximizing the expected utility of his real consumption of a mix of goods, as in F. Grauer-Robert H. Litzenberger-R. Stehle, will not convert forward his foreign investment returns and exchange receipts into his own currency—i.e. cover his exposure. He would do so, as Kenen's firm and B. A. Solnik's individuals are assumed to do, only if he desired to consume his own country's goods exclusively. Instead, he will seek to reduce his real purchasing power risks, arising from exogenously specified relative price and exchange rate changes. This can be achieved by diversifying into risky forward contracts when the spot rate is not perfectly correlated with the real returns on stocks and bonds.

The equilibrium results developed in Section III are useful for confronting the empirical literature and the predictions of the *MT*, most of which seeks to explain why forward rates deviate from *IRP*. Three strands of explanations have been offered.

Our explanation of departures of forward rates from *IRP* is founded on the existence of risks and the associated costs of default on bonds and forward contracts. Others have offered similar suggestions but without rigorous proof. Stoll (1972) and Grubel have argued that default risk in arbitrage transactions will make the (hypothetical) pure arbitrage function inelastic. Robert Z. Aliber surmises that probably political risks (i.e., of restrictions on convertibility) are enough to do the trick: this is not the case, however, as we show below.

Other authors largely rely on specific market imperfections. Paul A. Einzig, followed by Sohmen and Canterbery argue, after Keynes, for institutional constraints on the availability of arbitrage funds which, in the language of the *MT*, cause the arbitrage net-demand curve for forward contracts to become inelastic with respect to the forward rate. Martin F. Prachowny and Jacob A. Frenkel identify as the culprits capital market imperfections, in the form of borrowing rates which rise in an unspecified (and unmeasurable) fashion with the volume of arbitrage; Kenen makes a similar assumption. Such accounts are basically unsatisfactory in the sense that they require further justification of why institutions behave in the asserted fashion and how markets are cleared.

The third scenario began with William H. Branson, who attributes departures from *IRP* to transactions costs which create a band within which deviations can occur. Frenkel and R. M. Levitch suggest that once transactions costs are accurately accounted for and removed, *IRP* remains robust. This robustness may be illusory, however. The data used in em-

pirical studies are typically the prices quoted in the interbank dealer market, where, in the short term and under normal conditions, risks of default on interbank borrowing and of suspensions of convertibility are usually absent. In fact, it is well known that bank foreign exchange traders normally use the *IRP* formula to design their quotes. But deviations occur also in this market during abnormal periods (viz., the pound in 1967 and the Deutsche mark in 1973), when default and/or convertibility risks might have been high.

## I. The Portfolio-Choice Model

We postulate the existence of two nations separated by currency borders. Investors of both countries have access to three types of securities: stocks, bonds, forward exchange contracts. The bonds of both countries as well as the forward contracts are subject to a risk of default; i.e., there is a nonzero probability that these obligations may not be fulfilled. We assume, however, that these risks of default are exogenous and independent of the investors' decisions; they may result, for instance, from the government of either country suspending convertibility and thereby preventing a number of individuals from honoring their promises.

Individuals maximize a two-period utility-of-consumption function which exhibits no complementarity over time. Each period utility is the sum of a negative definite quadratic form and a linear form:[2] $-c'Qc + q'c$ where $c$ is the consumption vector. One could show that after optimization with respect to the consumption menu, the derived utility is:

$$(1) \qquad -c_0^2 + R_0c_0 + \alpha[-\tilde{w}^2/\tilde{I} + \tilde{R}\tilde{w}/\tilde{I}]$$

[2] This utility function is justifiable only as an approximation covering an infinitely short consumption period.

where:

$c_0$ = period-0 nominal consumption expenditures, expressed in currency 1;
$\tilde{w}$ = period-1 nominal wealth also equal to period-1 nominal consumption;
$\alpha$ = coefficient of time preference;

and $R_0$, $\tilde{I}$, $\tilde{R}$ are appropriate price deflators.[3]

Consider now the individual's investment decision. His future wealth will be given by:

$$(2) \qquad \tilde{w} = z'\tilde{V} + \tilde{\epsilon}(F - \tilde{S})C$$

where:

$z$ = vector of fractional holdings of each stock and bond security;
$V$ = vector of future payoffs of securities, all translated into currency 1;
$C$ = forward contracts to convert currency 2 into currency 1;
$\tilde{S}$ = future spot rate of exchange;
$F$ = time-0 forward rate;
$\tilde{\epsilon}$ = the random fractional loss to the investor if the forward contract is defaulted on; $0 \leq \epsilon \leq 1$.

Taking into account the budget constraint:

$$(3) \qquad w_0 - c_0 = z'V_0,$$

where $V_0$ is the vector of time-0 market values of securities, one can optimize (1) after substitution of (2). The resulting optimality conditions are:

$$(4a) \qquad Gz + E[\tilde{\epsilon}(F - \tilde{S})\tilde{V}/\tilde{I}]C = E(\tilde{R}\tilde{V}/\tilde{I})/2 - (R_0 - 2w_0)V_0/2\alpha$$

$$(4b) \qquad E[\tilde{\epsilon}(F\text{-}\tilde{S})\tilde{V}'/\tilde{I}]z + E[\tilde{\epsilon}^2(F\text{-}\tilde{S})^2/\tilde{I}]C = E[\tilde{R}\tilde{\epsilon}(F\text{-}\tilde{S})/\tilde{I}]/2$$

[3] $\tilde{I} = (\tilde{p}_1'Q^{-1}\tilde{p}_1)/(p_o'Q^{-1}p_o)$, $\tilde{R} = \frac{1}{2}\tilde{p}_1'Q^{-1}q$, $R_o = \frac{1}{2}p_o'Q^{-1}q$ where $p_o$ and $\tilde{p}_1$ are the time-0 and time-1 vectors of prices expressed in currency 1.

where:

$$G = V_0 V_0'/\alpha + E(\tilde{V}\tilde{V}'/\tilde{I}).$$

The determinant of the linear system (4) is $|G|\Delta$ where:[4]

$$(5)\quad \Delta = E[\tilde{\epsilon}^2(F\text{-}\tilde{S})^2/\tilde{I}] - E[\tilde{\epsilon}(F\text{-}\tilde{S})\tilde{V}'/\tilde{I}]G^{-1}E[\tilde{\epsilon}(F\text{-}\tilde{S})\tilde{V}/\tilde{I}].$$

Assuming that no two sets of stock or bond securities are perfectly correlated ($|G|\neq 0$), two cases arise. If $\Delta=0$, which implies that forward contracts and some linear combination of stocks and bonds are perfect substitutes,[5] then $C$ is indeterminate. If, however, $\Delta>0$, then:

$$(6)\quad C = E[(\tilde{R} - 2\tilde{V}'z_0)\tilde{\epsilon}(F\text{-}\tilde{S})/\tilde{I}]/2\Delta$$

where $z_0$ is the value of $z$ obtainable from (4a) when $C=0$, and $(\tilde{R}-2V'z_0)/\tilde{I}$ can be interpreted as the investor's marginal utility at zero forward contracting.

## II. Analysis of the Demand for Forward Contracts

We now study the composition and the variations of the demand equation (6) under the assumption that there exists no value of $F$ for which $\Delta=0$. We shall refer to this as the case where there is no perfect arbitrage opportunity.

The demand (6) can be decomposed into the sum of two terms. The first one is:

$$(7)\quad C_s = E[\tilde{R}\tilde{\epsilon}(F\text{-}\tilde{S})/\tilde{I}]/2\Delta$$

and represents the demand for forward contracts which would exist if no alternative instruments (stocks or bonds) were available for investment ($z=0$); it is therefore essentially a demand which is motivated by the relative levels of the forward and future spot rates. In accordance with traditional forward-market theory, we call it the "speculative demand," although, in the language of financial theory, it does not constitute "speculation" (see Jack Hirshleifer) but simply an investment in currencies. The speculative demand is zero for:

$$(8)\quad F_s = E(\tilde{S}) + \frac{\operatorname{cov}(\tilde{S}, \tilde{R}\tilde{\epsilon}/\tilde{I})}{E(\tilde{R}\tilde{\epsilon}/\tilde{I})}$$

where it is increasing (Figure 1). The covariance of the future spot rate with future goods prices and the default risk moves the intercept of the speculative demand away from the expected spot rate, contrary to the *MT*s assumption.

The second term in (6):

$$(9)\quad C_h = -E[\tilde{\epsilon}(F\text{-}\tilde{S})\tilde{V}'/\tilde{I}]z_0/\Delta$$

is a component of demand connected with the individual's portfolio investment and could perhaps be called "hedging demand." Hedging here is tantamount to diversification which the investor undertakes to reduce his real risks. This label should, therefore, not be identified with the one used by Grubel, for instance. The purpose of this hedging demand is *not*, as it was in the traditional analysis, to prepare the ground for the riskless conversion of the bulk of the investor's foreign income into his home currency. Rather, by diversifying into forward contracts, the investor intends to reduce the variability of his future real income, wherever it may originate. We note also that the investment so hedged includes investment in bonds as well as stocks; we do not refer to a covered bond transaction per se as "arbitrage" but rather as a hedged investment in bonds.

We reserve the term arbitrage for a motive which appears in the denominator $\Delta$

[4] It is necessary that $\Delta\geq 0$ in order for the second order condition to be satisfied.

[5] One could prove that $\Delta = 0$ if and only if two conditions are fulfilled: (a) $\tilde{\epsilon}(F-\tilde{S})$ is equal to a linear combination of $\tilde{V}$ and (b) the same linear combination of $V_o$ is null. The latter condition holds for only one value of $F$, which we denote $F_p$. See Adler and Dumas.

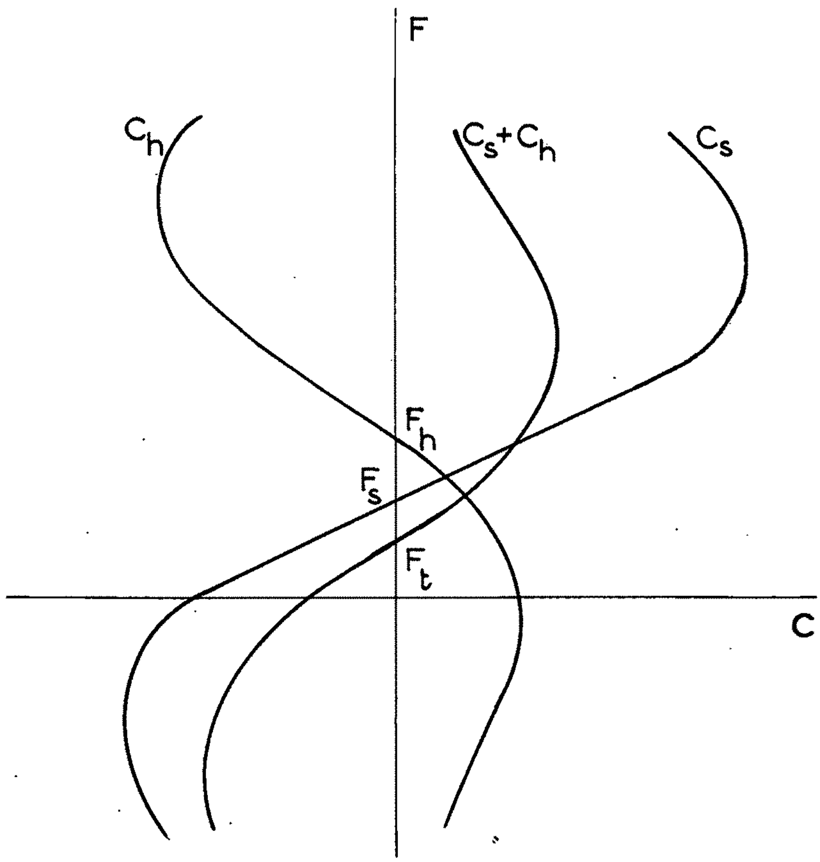



FIGURE 1. THE SPECULATIVE, HEDGING AND OVERALL DEMAND SCHEDULES WHEN THERE ARE NO PERFECT ARBITRAGE OPPORTUNITIES

of the demand. As the returns from forward contracts $\tilde{\epsilon}(F-\tilde{S})$ are progressively better approximated by a linear regression on the returns from stocks and bonds, $\Delta$ approaches zero for some value of $F$. In the neighborhood of that value, the investor takes advantage of the quasi arbitrage opportunity and the result is a magnification of his speculative and hedging demands.

Figure 1 makes two further points clear. The curvature of the $C_s$ curve seems to confirm traditional diagrammatic renditions of the speculation function. In the absence of perfect arbitrage opportunities the elasticity of the curve falls to zero at finite quantities of forward contracting. These curvature characteristics seem robust for well-behaved utility functions, other than the quadratic. But in those cases we have checked, it is impossible to decompose the total demand into the sum, $C_s$ and $C_h$, of two components. In all cases, however, the $C_s$ and $C_h$ functions cannot be observed separately.

We now consider briefly the case where there exist perfect arbitrage opportunities, i.e., where $\tilde{\epsilon}(F-\tilde{S})$ is equal to a linear combination of $V$, for some value $F_p$ of $F$. Then for $F=F_p$, if the same combination of $V_o$ is null, $\Delta(F_p)=0$ and $C$ becomes indeterminate. Furthermore $F_p$ is also a root of the numerator, so that for other values

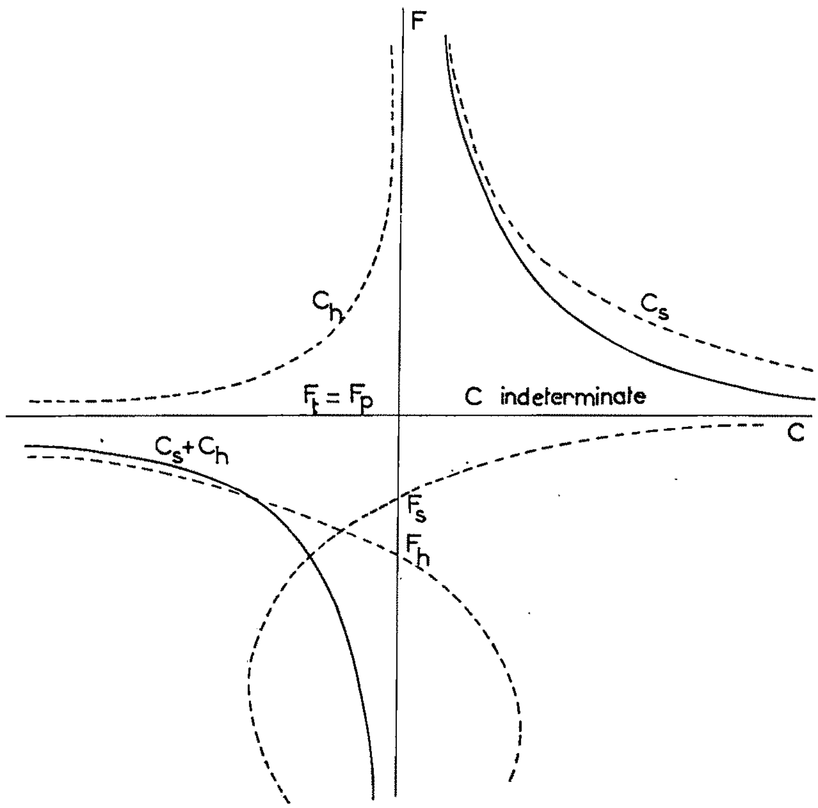



FIGURE 2. THE DEMAND SCHEDULE WHEN THERE EXIST PERFECT ARBITRAGE OPPORTUNITIES

of $F$, the demand reduces to a rectangular hyperbola (Figure 2).[6] The numerators and denominators of the speculative and hedging demands retain their original degrees, but infinite branches appear at $F=F_p$; this is an extreme illustration of the magnification effect of the arbitrage motive.

### III. The Equilibrium Forward Rate and the Predictions of the Modern Theory

As was mentioned in the introduction, the modern theory of forward exchange predicted that the equilibrium forward rate was always bracketed by the expected spot rate $E(\tilde{S})$ and the interest rate parity rate. To disprove this claim on the basis of the present model, we now consider the relationship between these three rates in the special case where the bonds are default free in both countries with interest rates denoted $r_1$ and $r_2$.

At equilibrium, the sum of all forward contracts is null: $\sum_k C_k = 0$, where $k$ subscripts represent the investors. Assuming that $\bar{I}$, $\bar{V}$, $\tilde{\epsilon}$ and $\tilde{S}$ are identical for all investors, we can apply this condition to (6) to get the equilibrium rate:

$$F = E(\tilde{S}) + \frac{\text{cov}\,(\tilde{S}, \tilde{M}\tilde{\epsilon})}{E(\tilde{M}\tilde{\epsilon})} \tag{12}$$

where $\tilde{M} = (\sum_k \tilde{R}_k - 2e'V)/\bar{I}$ is the inves-

[6] This hyperbola is somewhat of an oddity; it would not reflect rational behavior since, in the presence of a perfect arbitrage opportunity, the investor could instead reach infinite wealth by entering forward contracts in infinite amounts. The explanation is that, for large values of wealth and consumption, the quadratic approximation to the utility function (footnote 2) is no longer valid.

TABLE 1

| | | | cov $(\tilde{S}, \tilde{M})$ | |
|---|---|---|---|---|
| | | | $\leq 0$ | $\geq 0$ |
| cov $(\tilde{S}, \tilde{M}\tilde{\epsilon})$ | $\geq 0$ | $K \geq 0$ | $IRP \leq E(\tilde{S})$<br>$F \geq E(\tilde{S})$<br>$F \geq IRP$ | $IRP \geq E(\tilde{S})$<br>$F \geq E(\tilde{S})$<br>$F \geq IRP$ |
| | | $K \leq 0$ | Impossible | $IRP \geq E(\tilde{S})$<br>$F \geq E(\tilde{S})$<br>$F \leq IRP$ |
| | $\leq 0$ | $K \geq 0$ | $IRP \leq E(\tilde{S})$<br>$F \leq E(\tilde{S})$<br>$F \geq IRP$ | Impossible |
| | | $K \leq 0$ | $IRP \leq E(\tilde{S})$<br>$F \leq E(\tilde{S})$<br>$F \leq IRP$ | $IRP \geq E(\tilde{S})$<br>$F \leq E(\tilde{S})$<br>$F \leq IRP$ |

$$K = \frac{\text{cov}\,(\tilde{S}, \tilde{M}\tilde{\epsilon})}{E(\tilde{M}\tilde{\epsilon})} - \frac{\text{cov}\,(\tilde{S}, \tilde{M})}{E(\tilde{M})}$$

tors' aggregate marginal utility of wealth and $e = \sum_k z_{ok}$ is a vector containing for each security a one or a zero depending on whether this security's net supply is positive or null. In a similar fashion, one could derive the interest rate relationship from the optimality conditions (4):

$$\text{(13)} \qquad \frac{1 + r_1}{1 + r_2} = E(\tilde{S}) + \frac{\text{cov}\,(\tilde{S}, \tilde{M})}{E(\tilde{M})};$$

hence:

$$\text{(14)} \qquad F - \frac{1 + r_1}{1 + r_2} = \frac{\text{cov}\,(\tilde{S}, \tilde{M}\tilde{\epsilon})}{E(\tilde{M}\tilde{\epsilon})} - \frac{\text{cov}\,(\tilde{S}, \tilde{M})}{E(\tilde{M})}.$$

The default risk on forward contracts is, in this particular case, entirely responsible for the deviation from *IRP*. However, *contra* Aliber, if $\tilde{\epsilon}$ is independent of $\tilde{S}$ and $\tilde{M}$, the right-hand side of (14) is zero and *IRP* prevails even if $\tilde{\epsilon} < 1$. Of the six possible relations, outlined in Table 1, between $E(\tilde{S})$, $\tilde{F}$, $(1+r_1)/(1+r_2)$, only two, those emphasized by heavy outlines, correspond to the pattern predicted by the Modern Theory.

## IV. Conclusion

The main results of this paper have been summarized in the introduction. One policy implication, however, can be brought forth: under those conditions which make bonds and forward contracts perfect substitutes and which therefore also produce *IRP*, *it is impossible for the government successfully to affect the volume of short-term capital flows by intervening in the forward exchange market.* In this case, the forward exchange trading decisions are indeterminate for everyone and so, of course, is the equilibrium volume of forward contracting. The *MT* allows the speculative demand for foreign exchange, by intersecting the perfectly elastic arbitrage schedule, to determine the equilibrium volume of transactions and therefore implies that by "counter-speculation" the government can affect this volume along with the reverse arbitrage capital flows. But when forward

contracting is an arbitrary decision and forward contracts a redundant security, it is obviously impossible to specify or predict how much of them will be traded. Our theory bears this out.

## REFERENCES


M. **Adler** and B. **Dumas**, "Default Risk and the Demand for Forward Exchange," Columbia Univ., mimeo, Aug. 1975.

R. Z. **Aliber**, "The Interest Rate Parity Theorem: A Reinterpretation," *J. Polit. Econ.*, Nov.-Dec. 1973, *81*, 1451–59.

W. H. **Branson**, "The Minimum Covered Interest Differential Needed for International Arbitrage Activity," *J. Polit. Econ.*, Nov.-Dec. 1969, *77*, 1028–35.

E. R. **Canterbery**, "Foreign Exchange, Capital Flows and Monetary Policy," Princeton Studies in International Finance Section, Dept. of Economics, Princeton Univ. 1965.

P. A. **Einzig**, *A Dynamic Theory of Forward Exchange*, London 1961.

J. A. **Frenkel**, "Elasticities and the Interest Parity Theory," *J. Polit. Econ.*, May-June 1973, *81*, 741–47.

——— and R. M. **Levitch**, "Covered Interest Arbitrage: Unexploited Profits?" *J. Polit. Econ.*, Apr. 1975, 325–38.

F. **Grauer**, R. H. **Litzenberger** and R. **Stehle**, "Sharing Rules and Equilibrium in an International Capital Market under Uncertainty," Graduate School of Business, Stanford Univ., Feb. 1975, res. pap. 245.

H. G. **Grubel**, *Forward Exchange, Speculation and the International Flow of Capital*, California 1966.

D. **Heckerman**, "On the Effects of Exchange Risk," *J. Int. Econ.*, Nov. 1973, *3*, 379–88.

J. **Hirshleifer**, "Speculation and Equilibrium: Information, Risk, and Markets," *Quart. J. Econ.*, Nov. 1975, *89*.

P. B. **Kenen**, "Trade, Speculation and the Forward Exchange Rate," in R. E. Baldwin, et al., ed., *Trade Growth and the Balance of Payments*, Chicago 1965, 143–69.

J. M. **Keynes**, *A Tract on Monetary Reform*, London 1923.

M. F. **Prachowny**, "A Note on Interest Rate Parity and the Supply of Arbitrage Funds," *J. Polit. Econ.*, May-June 1970, *78*, 540–45.

D. **Schilling**, "Forward Exchange and Currency Position," *J. Fin.*, Dec. 1969, *24*, 875–86.

E. **Sohmen**, *Flexible Exchange Rates*, Chicago 1969.

B. H. **Solnik**, "Equilibrium in an International Capital Market," *J. Econ. Theory*, Apr. 1974.

J. L. **Stein**, "The Nature and Efficiency of the Foreign Exchange Market," Essays in Int. Finance 40, Princeton Univ. 1962.

H. R. **Stoll**, "An Empirical Study of the Forward Exchange Market under Fixed and Flexible Exchange Rate Systems," *Can. J. Econ.*, Feb. 1968, *1*.

———, "Causes of Deviation from Interest Rate Parity," *J. Money, Credit, Banking*, Feb. 1972, *4*, 113–17.

S. C. **Tsiang**, "The Theory of Forward Exchange and the Effects of Government Intervention in the Forward Exchange Market," *IMF Staff Papers*, Apr. 1959, 75–106.

APPLICATION OF OPTIMAL CONTROL TO PROBLEMS OF ECONOMIC STABILIZATION

# Control Methods for Macroeconomic Policy Analysis

*By* GREGORY C. CHOW*

## I. Why Optimal Stochastic Control?

The use of econometric models for the quantitative analysis of macroeconomic policy is now generally accepted. A great deal of research in the last several years has been devoted to the development and application of optimal stochastic control techniques in policy analysis. This development appears to be natural and necessary. Using an econometric model, one can make projections of the key economic variables in future periods given a set of proposed future values for the policy variables or instruments. One may then examine the nature of these projections in order to evaluate a policy proposal. This approach to policy analysis is deficient for two reasons.

First, the dynamic response of the economic variables to a particular course assigned to the policy variables is complicated and unpredictable. This makes the selection of policy by trial and error extremely inefficient. There is a need to specify a loss function of the key economic variables and to minimize its value with respect to the policy instruments. The specification of an objective or loss function and the derivation of a policy solution by optimization is an essential feature of optimal control. It is much more efficient than the method of trial and error. The objective function is also useful for the evaluation of other policy proposals than the optimum policies.

The second, and perhaps more important, reason is due to the uncertainty of economic projections. Given the proposed time paths for the policy variables, one cannot rely on an econometric model to make perfect predictions of the important economic variables. This uncertainty not only makes it difficult to evaluate a given policy path, but it makes the evaluation of such a path unrealistic and irrelevant.

The former difficulty can be resolved by stochastic simulations which incorporate the random elements in the econometric model in making projections; the means and variances of the future paths can then be calculated. Because of uncertainty in the economy, a policy maker often will not adhere to a fixed plan irrespective of future developments. Future decisions will be made on the basis of future observations of the economy. Hence, it is usually unrealistic and irrelevant to evaluate the consequences of a preassigned sequence of policy actions. A more realistic policy takes the form of a reaction function, or a feedback control equation; it is a rule to determine the future values of the policy variables according to future economic observations. By the technique of optimal stochastic control, the solution to a multiperiod planning problem under

* Economic Research Program, Princeton University. Financial support from the National Science Foundation is gratefully acknowledged. This paper was typed by Alice Furth.

uncertainty is given by a set of feedback control equations. The solution in feedback form is important from both the descriptive and the normative points of view.

We can therefore conclude that in policy analysis it is necessary to optimize, and optimization has to be performed in a stochastic setting to yield optimal feedback control equations. Deterministic control theory ignores uncertainty in the econometric model and yields a preassigned future path for the policy variables as a solution. This deterministic solution path, nevertheless, has one interesting characteristic. If the econometric model is linear with an additive random disturbance, if its parameters are known for certain, and if the objective is to minimize the expected value of a quadratic loss function for T periods, then according to a theorem due to Herbert A. Simon and Henri Thiel, the optimum solution for the first period is identical with that of the above deterministic path. This result is known as the first-period certainty equivalence principle. Imagine how this principle can be applied to evaluate the outcome of an optimal sequential plan for a stochastic control problem with twenty periods as the planning horizon. A twenty-period deterministic control problem is solved to obtain $x_1$, the optimum setting of the vector of policy variables for the first period. Given $x_1$, we generate $y_1$, the vector of endogenous variables in period 1, using the econometric model and a random drawing for the vector $u_1$ of random disturbances. Having generated $y_1$ stochastically, we have the initial condition to solve a nineteen-period deterministic control problem to obtain the optimal $x_2$ for the second period. We then generate $y_2$ from $x_2$ and a random drawing of the vector $u_2$ of random disturbances for the second period, using the econometric model. An eighteen-period deterministic control problem is then solved to obtain $x_3$, and so forth until $x_{20}$ and $y_{20}$ are obtained. All this will give one observation of a sequence of the $x$'s and the $y$'s for twenty periods. The process can be repeated $n$ times. Each observation provides a value for the loss function, and the expected loss can be estimated by averaging over the $n$ observations. This would be a laborious and costly procedure. By contrast, using the technique of stochastic control to be described in the next section, we derive a set of feedback control equations which determine the optimal values for the future policy variables in terms of the values of future economic observations. The minimum expected loss associated with the optimal policy can be calculated by a simple formula. Furthermore, the means, variances and covariances of the future paths of all economic variables can be obtained analytically without resort to stochastic simulations. They contain much useful information concerning the dynamic performance of the economy under the rules of optimal stochastic control.

## II. Techniques of Stochastic Control

Before presenting some important applications of stochastic control to the analysis of macroeconomic policy, I will survey briefly the available techniques. Of most importance is the derivation of the optimal feedback control equations and the associated expected welfare loss. Solutions are now available for linear as well as nonlinear econometric models, as described in Chow (1975, 1976a, 1976b). The techniques can also be classified according to the treatment of uncertainty. Beginning with the case of complete certainty where all model parameters are assumed to be known and all random disturbances are set equal to their mean value zero, one can first introduce random disturbances and secondly allow for uncer-

tainty in the model parameters. We will postpone till the end of this paper the treatment of a third kind of uncertainty, namely in the specification of the model itself. For the control of a linear model with a vector of random disturbances but known parameters, assuming a quadratic loss function, we can find the optimal solution as a set of linear feedback control equations. When uncertainty is introduced into the parameters of a linear model, and when either the parameters or the disturbances in a nonlinear model are treated as uncertain, the solutions available are only approximately optimal. Stochastic control theory is still useful for the calculation of the expected loss associated with the approximately optimal policies and the derivation of the important dynamic characteristics of the economy under control.

For expository purposes in this section, I will illustrate the solution to the optimal control of a linear system with unknown parameters using a quadratic loss function. The system is written as

$$(1) \quad y_t = A_t y_{t-1} + C_t x_t + b_t + u_t$$

where $y_t$ is a vector of endogenous variables at time t, $x_t$ is a vector of policy variables at time t, $b_t$ is a vector combining the effects of all exogenous variables not subject to control, the matrices $A_t$, $C_t$, and $b_t$ consist of unknown parameters whose probability distribution is assumed to be given, and $u_t$ is a vector of random disturbances having mean 0, covariance matrix $V$, and being serially uncorrelated. Endogenous variables and policy variables with higher order lags can be eliminated by defining new endogenous variables so as to retain the form (1) of a system of first order linear stochastic difference equations in which only the current control variables $x_t$ appear. We can include the policy variables in the vector $y_t$ so that $x_t$ need not be an argument of the loss function. After suitable transformations, the model (1) can also be used to deal with serially correlated random disturbances. These technical details are explained in Chow (1975). The loss function for a T period control problem is

$$(2) \quad W = \sum_{t=1}^{T} (y_t - a_t)' K_t (y_t - a_t)$$

where $a_t$ is a vector of targets for the variables $y_t$ and $K_t$ is a diagonal matrix giving the relative penalties for the squared derivations of various variables from their targets.

To ease the exposition without effecting the essential argument, let me assume that $a_t = 0$ and $b_t = 0$ for all t. The problem is to minimize the expected value of the loss function for T periods by choosing a strategy for $x_1, x_2, \ldots, x_T$. The control variables will be selected sequentially, the vector $x_t$ for each period being determined only after the up-to-date information is available. This information consists mainly of $y_{t-1}$ which includes the observations of all past endogenous variables and policy variables affecting the current endogenous variables at time t. Using the method of dynamic programming, we first solve the problem for the last period T by minimizing

$$\begin{aligned}(3) \quad V_T = E_{T-1}(y'_T K_T y_T) &= E_{T-1}(y'_T H_T y_T) \\ &= y'_{T-1}(E_{T-1} A'_T H_T A_T) y_{T-1} \\ &\quad + x'_T(E_{T-1} C'_T H_T C_T) x_T \\ &\quad + 2x'_T(E_{T-1} C'_T H_T A_T) y_{T-1} \\ &\quad + E_{T-1} u'_T H_T u_T\end{aligned}$$

where we have used (1) to substitute for $y_T$, taken mathematical expectation $E_{T-1}$ conditioned on the data available at the end of period T−1, assumed $a_t = 0$ and $b_t = 0$, and defined $K_T = H_T$ for future convenience. Minimizing (3) with respect to the vector $x_T$, one obtains the optimal feedback control equation for the last period,

$$(4)\quad \hat{x}_T = -(E_{T-1}C'_T H_T C_T)^{-1} \cdot (E_{T-1}C'_T H_T A_T) y_{T-1} = G_T y_{T-1}.$$

(Without the assumptions $a_t=0$ and $b_t=0$, the solution (4) would contain an intercept $g_T$.) The minimum expected loss for the last period associated with the optimum policy is obtained by substituting (4) for $x_T$ into (3):

$$(5)\quad \hat{V}_T = y'_{T-1}(EA'_T H_T A_T) y_{T-1} - y'_{T-1}(EA'_T H_T C_T)(EC'_T H_T C_T)^{-1} \cdot (EC'_T H_T A_T) y_{T-1} + Eu'_T H_T u_T.$$

To solve the problem for the last two periods, we apply the principle of optimality in dynamic programming to minimize with respect to $x_{T-1}$ the expression

$$(6)\quad V_{T-1} = E_{T-2}(y'_{T-1}K_{T-1}y_{T-1} + \hat{V}_T).$$

The rationale for (6) is that whatever $x_{T-1}$ is chosen, which will affect the outcome $y_{T-1}$, we shall choose the optimum policy $\hat{x}_T$ according to the feedback control equation (4) to yield the minimum expected loss $\hat{V}_T$. The optimal policy for both periods can thus be obtained by minimizing (6) with respect to $x_{T-1}$, given that $x_T$ will be optimally chosen. Substituting (5) into (6) and letting

$$(7)\quad H_{T-1} = K_{T-1} + (EA'_T H_T A_T) - (EA'_T H_T C_T)(EC'_T H_T C_T)^{-1} \cdot (EC'_T H_T A_T),$$

we rewrite (6) as

$$(8)\quad V_{T-1} = E_{T-2}(y'_{T-1}H_{T-1}y_{T-1}) + Eu'_T H_T u_T.$$

Note that, except for a constant, (8) has the same form as (3) with the subscript T−1 replacing T. Therefore, by the same argument, the optimum $\hat{x}_{T-1}$ will be given by equation (4) with T−1 replacing T; the minimum total expected loss $\hat{V}_{T-1}$ for periods T−1 *and* T will then be found as in (5). The process continues until $\hat{x}_1$ and $\hat{V}_1$ are obtained. $\hat{V}_1$ is the minimum total expected loss for all T periods. It can be computed by equation (5) with the subscript 1 replacing T, where $H_1$ is obtained by solving the difference equation (7) for $H_t$ backward in time. In these calculations, the approximation has been made that all expectations involving the unknown parameters $A_t$ and $C_t$ in the future are based only on data available at the beginning of period 1. Otherwise, the derivation will be more complicated because the expectations in equation (4) are functions of the data $y_{T-1}$, $y_{T-2}$, etc. We have just demonstrated how the optimal feedback control policies and the associated expected loss for all periods can be obtained analytically. The above derivations can be extended to deal with nonlinear econometric models with unknown parameters, as described in Chow (1976a).

## III. Applications to Macroeconomic Policy Analysis

Having elaborated on the two major concepts in optimal stochastic control, namely, the minimum expected loss and the feedback control equations, we will illustrate their applications to the analysis and formulation of macroeconomic policies.

Using the minimum expected loss, one can compare various policy proposals as described by different feedback control equations or reaction functions. The candidates for comparison include such control equations as the ones prescribing the policy variables to grow at constant percentage rates, reaction functions estimated from time series data as descriptive of the historical decision making process, and policies which rely exclusively on a subset of active instruments while keeping the re-

maining passive instruments to grow at some constant rates in order to study the relative effectiveness of fiscal and monetary policies. The minimization of expected loss can also be used to measure the trade-off possibilities for inflation and unemployment implicit in an econometric model. Without using the techniques of optimal control, one can experiment with different policies using an econometric model and observe the resulting rates of inflation and unemployment, hoping to trace out the trade-off possibilities. However, this approach is defective because the inflation-unemployment combinations generated by the experiments may not be the best possible. By varying systematically the parameters of the loss function in an optimal control formulation, one can trace out the best combinations of inflation and unemployment that can be achieved according to a given econometric model. Finally, the minimum expected loss can be used to measure the value of information and the costs of delays. The amount of information can be described by the variances of the estimated parameters and of observation errors; the variance-covariance structure, when incorporated into the analysis, will affect the minimum expected loss. Delays in information and in carrying out the decisions can also be modeled in the framework of stochastic control and the resulting minimum expected losses can be compared, Chow (1975, pp. 182ff, 202).

Secondly, the feedback control equations are extremely useful for policy analysis. When combined with the other econometric equations, they become a part of a stochastic model whose dynamic properties can be studied by analytical techniques. For the analysis of any policy, optimal or not, one would like to ascertain the means, variances and covariances of the important time series generated by an econometric model given the policy. Spectral properties of these time series can also be deduced analytically, as described in Chow (1975). Since automatic stabilizers can be specified by feedback equations, one can study the dynamic properties of an economy subject to different stabilization schemes. One can even optimize with respect to the parameters in the automatic stabilization schemes by the techniques of stochastic control for nonlinear systems in order to find a good scheme. Various dynamic policy multipliers, including the impact, delayed, interim, and long-run multipliers, can be calculated from a nonlinear econometric model after appropriate linearizations; they are by-products of the optimal control calculations for nonlinear models which I have proposed. The tools of stochastic control can also be used to compare econometric models, in terms of the optimal feedback control equations which they imply, of the degrees of stability under the optimal stabilization policies, and of the sensitivities of the control solution to changes in the parameters of the models or the loss functions. In short, the techniques of stochastic control can be used to deduce the policy implications of econometric models systematically and efficiently.

Let me conclude this paper by a proposal to deal with the third kind of uncertainty in econometric models, the uncertainty associated with the specification of the model itself. The decision maker, when faced with a serious choice between two or three competing models, is advised to calculate the optimal policies based on these models, and to examine how these and other proposed policies perform under the assumptions of the different models. A payoff matrix should be used, with different columns corresponding to the different models and different rows corresponding to the different policies to be examined, each element in the matrix being the expected loss asso-

ciated with a combination of model and policy. Such an analysis might uncover policies that are superior to the inactive policies for all models. This finding would be extremely important. If no such active policies could be uncovered which dominate the inactive policies for all the seriously competing models, much useful information would have been obtained on the degrees of differences in these econometric models, the needed areas of econometric research for macroeconomic decision making, and the empirical basis for current policy recommendations or decisions. A Bayesian would assign probabilities to the competing econometric models and choose that strategy which minimizes the expected value of the entry in the above payoff matrix. To any rational economic decision maker, Bayesian or not, the techniques of optimal stochastic control are important not because existing econometric models are nearly perfect but because how imperfect the models are, what imperfections matter for policy analysis, and what areas require most research can be ascertained efficiently only by the use of these techniques.

REFERENCES

G. C. Chow, *Analysis and Control of Dynamic Economic Systems,* New York 1975.

———, "The Control of Nonlinear Econometric Systems with Unknown Parameters," *Econometrica,* May 1976a, *44.*

———, "An Approach to the Feedback Control of Nonlinear Econometric Systems," *Annals of Economic and Social Measurement,* Spring 1976b, *5.*

H. A. Simon, "Dynamic Programming Under Uncertainty with a Quadratic Criterion Function," *Econometrica,* Jan. 1956, *24,* 74–81.

H. Theil, "A Note on Certainty Equivalence in Dynamic Planning," *Econometrica,* Apr. 1957, *25,* 346–49.

# Some Stabilization Problems of 1971-75, with an Application of Optimal Control Algorithms

*By* ALBERT ANDO AND CARL PALASH*

This paper attempts to analyze the relative contribution of various exogenous factors and of monetary and fiscal policies to the inflation and recession of 1973–75, using the MIT-PENN-SSRC (*MPS*) model and an optimal control algorithm applied to this model. Our conclusion is that while the rate of inflation in this period could not have been reduced much by aggregate monetary and fiscal policies, it was possible to have maintained a much lower rate of unemployment without worsening the rate of inflation by appropriate monetary and fiscal policies. We also suggest a possible reason why policy makers may have considered the very restrictive policies of 1974 appropriate at that time.

When the basic cause of inflation is excess final demand, a necessary ingredient of any policy to control inflation is some measure to reduce final demand. But, to the extent that the inflation is caused by factors other than excess demand, any policy measure to reduce final demand is unlikely to be effective in controlling inflation. In order to make this proposition more precise and quantitative, we need to identify the relative contribution of several factors to the inflation of 1974–75 period.

When the *MPS* model, which is used for our analysis here, or any other fairly large scale, economy-wide econometric model is simulated with residuals of all equations fed back into their respective equations, the model reproduces the historical path of all endogenous variables identically. In such a simulation, all exogenous shocks to the economy, such as the sudden increases of the price of oil, are incorporated either in values of exogenous variables or in values of residuals. Taking advantage of this feature, we can resimulate the model beginning in the middle of the 1960's with appropriate changes in monetary and fiscal policies so that the unemployment rate is always kept above 5 percent with the average of around 7 percent throughout the later 1960's and 1970's. Under such a policy, we enter the 1970's with the rate of inflation close to zero, and we can interpret the rate of inflation generated in such a simulation for 1974–75 period as that due to factors other than excess demand. In terms of the deflator of domestic, private, nonfarm product, the rate of inflation averages 8.5 percent in 1974, with the maximum of 10 percent reached in the second quarter. The actual, historical rate of inflation in terms of the same index is 12.8 percent for 1974. These figures suggest that at least two-thirds of the inflation in 1974

* Dept. of Economics, University of Pennsylvania and the Federal Reserve Bank of New York, respectively. This is a summary of a longer paper on the same subject which is scheduled to appear in *Frontiers of Quantitative Economics,* Vol. III, edited by Michael Intriligator, North-Holland Publishing Co., Amsterdam 1976, where data and analytical results that support statements made here are presented. Ando's contribution to this paper has been partially supported by the National Science Foundation under grant GS-32383X. Palash's participation in the preparation of this paper does not imply that views expressed here reflect those of the Federal Reserve Bank of New York or its staff.

was due to factors other than excess final demands. Further detailed analysis reveals that, in addition to such factors as the sudden increase of the oil price, the rise in food prices, and the devaluation of the dollar, two major contributors to inflation were the loss of productivity per manhour and quite significant increases in the markup in final good prices over the minimized average cost.

Against this background, both exogenous shocks given to the *U.S.* economy and monetary and fiscal policy postures adopted by the federal government were unusually deflationary. Exogenous increases of raw materials prices have basically the same effect on the economy as imposition of additional excise taxes, and the order of magnitude was extraordinarily large. And because the decline in final demand and output was accompanied by increases in prices, the standard automatic stabilizers worked in a perverse manner, leading to the full employment surplus in the federal budget of some $20 to $30 billion in the fourth quarter of 1974. On the monetary side, the short-term money market rates of interest were raised to unprecedented levels during the spring and summer of 1974. Although the actual rate of inflation was very high, there is good reason to believe that the expected rate of change of prices was much less than the actual rate of inflation. Furthermore, there is every indication that the market-required risk premium in equity rose very substantially during this period. The net result was that the long term required rate of return appears to have risen rapidly in 1974. Thus, the severe recession of 1974–75 can be fully explained by exogenous shocks and monetary and fiscal policy actions of this period. Moreover, there is some indication that the severity and rapidity of the decline in final demand, the reasons for which were not fully comprehended by the government policy makers and businessmen alike, contributed substantially to the loss of productivity per manhour and to the increased markup, and hence to inflation.

Once these points are clarified, we can see why a much less restrictive set of monetary and fiscal policies in 1974 would probably have resulted in much less severe loss of output without noticeable worsening of inflation. Using the *MPS* model, we show that this conclusion is indeed highly plausible, and the rate of unemployment in 1974–75 period could have been contained at no more than 6.5 percent or so without any worsening of inflation.

The monetary and fiscal authorities in this period clearly did not explicitly aim at creating recession, but they concentrated their attention on inflation and attempted to reduce the rate of inflation by restricting final demands. Since the inflation was not primarily caused by excess final demand, this policy design resulted in a sharp reduction of demand without much moderation of inflation. This point is most clearly brought out by an application of optimal control techniques to the analysis of policies followed in this period.

In applying optimal control techniques to the problem at hand, the most delicate problem is the specification of the objective function to be maximized. The common approach is to use a function whose first and second derivatives with respect to target variables (with appropriate signs) are all negative. This procedure works more or less satisfactorily under normal conditons. However, when one or more target variables cannot be reduced below a certain level due to external factors, the maximization of such an objective function subject to an econometric model without explicit allowance for this constraint produces a set of policies that are "unreasonable." This is because, in these circumstances, the negative second

derivative of the objective function with respect to the constrained target variable makes the implicit weight given to this variable much greater than intended.

We have shown that the maximization of a seemingly reasonable objective function without explicit allowance for the minimum inflation rate due to external factors in 1973–75 period does indeed generate a set of policies very similar to the ones actually followed. When the implicit relative weights attached to inflation and unemployment in the objective function generating this result are calculated, the weight for inflation turns out to be so high that the weight given to unemployment is almost insignificant. This situation can be easily remedied once we allow for the inflation rate that is due to external factors in the objective function.

The problem of unintentionally assigning "unreasonable" weights to target variables as a result of incorrect analysis of economic conditions facing policy makers was handled explicitly in this paper using an optimal control framework. It is quite plausible that, in formulating monetary and fiscal policies in much less formal manner, policy makers can nevertheless be trapped into a similar situation, leading to an imposition of inappropriate set of policies on the economy. We suggest that this is what has happened in 1973–75 period.

# On the Use of Feedback Control in the Design of Aggregate Monetary Policy

*By* J. H. Kalchbrenner and P. A. Tinsley*

## I. A Feedback Strategy for Monetary Policy

Periodic feedback revision in policy plans to incorporate recent measurements of economic activity is perhaps the minimal strategy response to uncertainty. This paper illustrates an evolution of feedback revisions in monetary policy by an optimal control exercise with the MIT-PENN-SSRC (*MPS*) quarterly model over the turbulent eight-quarter interval beginning in mid-1973. The purpose of the exercise is to trace the impact of measured historical dislocations of the economy on the *ex ante* expectations of a feedback strategy, and contrast the result with two alternative strategies: an open-loop strategy without feedbacking, and the optimal "prescience" strategy based on perfect foresight.

The design of the feedback exercise is to revise *ex ante* monetary policy at six-month intervals: t=1973–III, 1974–I, 1974–III, and 1975–I. At the beginning of each recontract quarter, the hypothesized planners select an eight-quarter path for nonborrowed reserves that minimizes the expected loss for the next eight quarters conditioned on their current information set $E\{L_t|t-1\}$.

Policy loss for this rolling horizon procedure was represented by an eight-quarter sum of asymmetric components

$$(1)\quad L_t = \sum_{j=t}^{t+7} [2(U_j > 4.8)^2 + (G_j^p > 2.5)^2 + 5(|\Delta R_j| > 1.5)^2 + 10^{-4}(M_1 - M_1^*)_j^2],$$

where the first term penalizes unemployment exceeding an approximation of the "natural rate" of the *MPS* quarterly model (and zero otherwise); similarly, the second term is the square of the annual rate of change in the implicit deflator of private business output that exceeds 2.5; the third term inflicts a sizeable penalty on quarterly swings in the 90-day Treasury bill rate that exceed 150 basis points; and the last term imposes a mild penalty on deviations from the 5.1 percent growth path in $M_1$ that is approximately consistent with the long-term price objective in (1).

To avoid overstating the amount of information that was historically available at each recontract quarter, the hypothetical planners were provided with *ex ante* forecasts of 17 exogenous variables and selected "add factor" forecasts of the model residuals (i.e., the random disturbances of the structural equations) that were based on historical assumptions by the *FRB* model group in the quarter prior

* Adviser and Chief (Special Studies), Division of Research and Statistics, Federal Reserve Board. We are indebted to L. Hanson, A. Havenner, and especially J. Berry for superb programming. We are also grateful to J. Enzler for discussion of historical forecasting procedures. An introduction to control techniques by analogy to statistical estimation and a more detailed analysis of Federal Reserve Board (*FRB*) staff procedures from the perspective of control theory can be found in Kalchbrenner and Tinsley (1975). The views expressed herein are solely those of the authors and do not necessarily represent the views of the Board of Governors of the Federal Reserve System.

TABLE 1—EVOLUTION OF FEEDBACK FORECASTS AND THE IMPACT OF HISTORICAL DISTURBANCES

| | *Ex Ante* Expectations | | | | *Ex Post* Results | | |
|---|---|---|---|---|---|---|---|
| | Feedback | | | | | | |
| | 73-III | 74-I | 74-III | 75-I | Feedback | Pseudo History | Prescience (H=14) |
| | | | $G^p$ % Price Change (private output) | | | | |
| 73-III | 4.8 | | | | 5.7 | 5.7 | 5.7 |
| -IV | 5.9 | | | | 8.5 | 8.4 | 8.4 |
| 74- I | 6.0 | 5.2 | | | 12.6 | 12.6 | 12.6 |
| - II | 5.6 | 3.5 | | | 13.6 | 13.7 | 13.8 |
| -III | 3.3 | 5.4 | 6.9 | | 12.1 | 12.2 | 12.7 |
| -IV | 3.2 | 3.0 | 6.6 | | 12.5 | 12.7 | 13.5 |
| 75- I | 4.6 | 5.4 | 7.3 | 6.8 | 11.4 | 11.5 | 12.4 |
| - II | 2.8 | 4.7 | 5.9 | 5.3 | 3.8 | 3.7 | 4.5 |
| | | | $U$ Unemployment Rate | | | | |
| 73-III | 4.5 | | | | 4.6 | 4.6 | 4.7 |
| -IV | 4.5 | | | | 4.6 | 4.6 | 4.7 |
| 74- I | 4.6 | 4.7 | | | 5.0 | 5.0 | 5.1 |
| - II | 4.9 | 5.0 | | | 5.1 | 5.0 | 4.9 |
| -III | 5.2 | 5.2 | 5.6 | | 5.6 | 5.4 | 4.9 |
| -IV | 5.4 | 5.3 | 6.1 | | 6.7 | 6.4 | 5.5 |
| 75- I | 5.7 | 5.4 | 6.5 | 7.5 | 8.3 | 8.2 | 6.9 |
| - II | 5.3 | 5.5 | 6.9 | 7.9 | 8.7 | 8.7 | 7.3 |
| | | | $G^m$ % $M_1$ Growth | | | | |
| 73-III | 3.4 | | | | 3.1 | 0.7 | −1.1 |
| -IV | 1.5 | | | | 2.5 | 7.7 | 7.0 |
| 74- I | 7.9 | 5.5 | | | 9.2 | 6.9 | 14.5 |
| - II | 3.5 | 1.6 | | | −0.9 | 6.3 | 20.5 |
| -III | 8.3 | 9.0 | 7.8 | | 8.5 | 1.4 | 5.4 |
| -IV | 5.5 | 10.5 | 10.9 | | 4.9 | 3.0 | −0.3 |
| 75- I | 0.6 | 3.9 | 2.4 | 7.0 | 9.2 | 4.7 | −4.3 |
| - II | 9.4 | 7.5 | 7.8 | 20.4 | 20.4 | 10.7 | 1.6 |

to the relevant recontract date. The selected exogenous variables include conditional forecasts of federal expenditures and tax rates in addition to projections of the price deflators of exports and imports. The set of implicit add factors was constructed by adjusting selected intercepts of the structural model until the forecasts of unemployment $U$ and inflation $G^p$ conditioned on both the *ex ante* exogenous variable projections and a given $M_1$ path coincided with the historical forecasts associated with that $M_1$ path. In all cases, the projections were adjusted for subsequent revisions in the data.

Although the aim of this exercise is to assess the value of a feedback strategy with a policy model in a realistic information con-

text, it is perhaps inevitable that the "optimal" feedback response will be compared with historical monetary policy. There are several differences in procedure that make a strict comparison invalid. First, the control instrument in the feedback exercise is nonborrowed reserves. Second, the information sets were based on historical files generated as staff background material and do not correspond to actual briefing material that is heavily modified by information from judgmental data banks. Third, the feedback exercise recontracts every six months whereas historical monetary policy is formally reassessed at monthly intervals. Fourth, the effective loss function and planning horizon of the planners is not known. Fifth, the implicit assertion that the exogenous variables and historical disturbances are independent of the monetary policy control settings is a debatable assumption. Finally, to preclude discussion of current policy, the *ex ante* feedback plans for quarters subsequent to 1975–II are not shown.

A portion of the results of the feedback exercise is listed in Table 1. The first four columns depict the evolution of *ex ante* expectations concerning the unemployment rate $U$ and the rate of growth in the price level $G^p$ and money $G^m$. The first column, for example, lists the forecasts that are conditioned on both the 73–III information set and the nonborrowed reserves path that minimized the eight-quarter loss function (1) initialized in 1973–III. The quarterly model was then simulated for two quarters along the "optimal" nonborrowed reserves path where the *ex ante* information set is replaced by the historical exogenous variables and stochastic disturbances. The results of this two-quarter simulation are listed in the *ex post* feedback column of Table 1. These simulation results are then the initial conditions for the next optimization and simulation cycle. This eight-quarter optimization and two-quarter simulation cycle is repeated for each of the three remaining recontract periods.

For rough comparison, the *ex post* portion of Table 1 also lists (a) a *pseudo-historical* simulation where nonborrowed reserves was set on its historical path, and (b) a *prescience* solution that is the result of minimizing a 14-quarter horizon where the historical exogenous variables and disturbances were known for the first eight quarters and the projections for the remaining six quarters were conditioned on the information set of 1975–I.

Table 1 indicates the difficulties in this period of obtaining accurate medium-run forecasts of the arguments of the policy loss function and the apparent insensitivity of the feedback strategy to measurement of large forecast errors. If we (somewhat imprudently) characterize the monetary policy intentions of the planners by the *ex ante* $M_1$ projection rates, there is a modest initial response to a sequence of

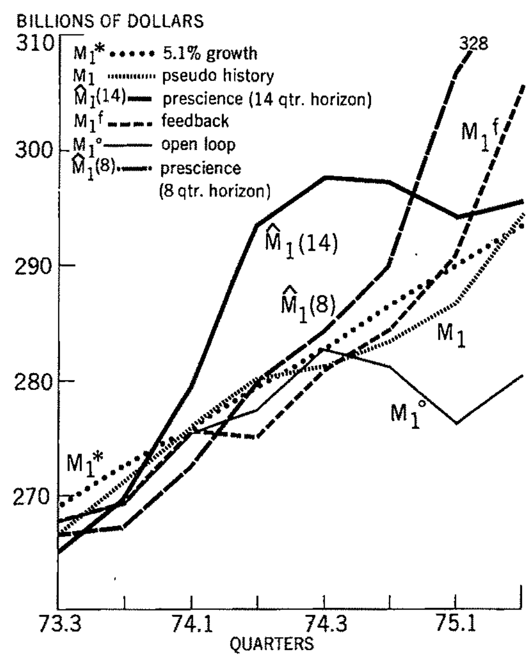



FIGURE 1: $M_1$ TRAJECTORIES FOR ALTERNATIVE STRATEGIES

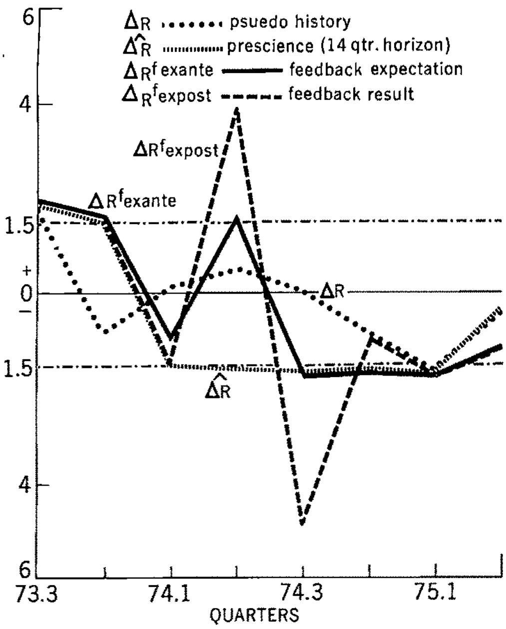



FIGURE 2: BILL RATE CHANGES PRODUCED BY ALTERNATIVE STRATEGIES

severe misses in the conditional forecasts of inflation. By the 74–III plan, it is evident that the rise in unemployment projected for the first half of 1975 is more severe than previously expected; however, *ex ante* planned expansion of $M_1$ in the fourth quarter is thwarted (compare *ex ante* and *ex post* $M_1$ growth in 74–IV). A pronounced departure in effective $M_1$ growth from the open loop plan (the 73–III column) is not accomplished until the first half of 1975.

Hindsight is suggestive but ultimately frustrating. As shown in Figure 1, the optimal prescience solution for $M_1$ (based on a 14-quarter horizon) is an interesting contrast to both the feedback $M_1^f$ and historical $M_1$ paths. The eight-quarter growth rate of the prescience solution is approximately equal to the growth of historical $M_1$ (and of the long-term desired path $M_1^*$, for that matter) but most of the expansion is dumped into the first half of 1974. The result of this growth rate pattern is a substantial easing of the unemployment rate in the last half of the period at the cost of a moderate rise in the rate of change of the implicit deflator.

The feedback strategy hits the accelerator hard about four quarters later than the prescience response. This exaggerated policy response is fairly typical of control solutions that assign all uncertainty to the stochastic disturbances of the final form equations. (See discussion in Kalchbrenner and Tinsley.) Whereas the amplitude of the feedback response would probably be moderated by allocating at least a portion of the uncertainty to the "slope" coefficients of the final form model, it is not obvious that more frequent filtering would substantially advance the timing of the feedback response.

The prescience solution based on an eight-quarter horizon $\hat{M}_1(8)$ is thrown in for contrast with $\hat{M}_1(14)$ to illustrate the impact of a short planning horizon. The relative oscillation of $\hat{M}_1(8)$ is an attempt to play off the differential policy response lags in inflation and unemployment against the short-horizon assumption that the world expires at the end of 1977–II.

To complete the record of state performance, the Treasury bill rate swings produced by alternative strategies are sketched in Figure 2. Note that interest rate smoothing was a binding constraint on the effective feedback response (e.g., the 74–III *ex ante* plan). The discrepancy between the *ex ante* expectations and *ex post* results of the feedback strategy for the bill rate and $M_1$ reflects the decision to stick to the preselected nonborrowed reserves path for six-month intervals; if deemed desirable, the *ex post* swings could be attenuated by more frequent feedbacking.

## II. A Brief Post-Mortem

Since, as a general rule, certainty equivalence strategies tend to be unusually sensitive to initial condition displacements, the cautious initial response of the feed-

TABLE 2—EXAMPLES OF RELATIVE FORECAST ERROR[a] EVOLUTION FOR GNP INFLATION ($g^p$) AND CIVILIAN UNEMPLOYMENT ($u$) RATES

| | Date | Fair | | | Wharton | | | "Consensus"[b] | |
|---|---|---|---|---|---|---|---|---|---|
| | $\tau$ \ t | 74.01 | 74.07 | 75.01 | 73.12 | 74.03 | 74.12 | 73.12 | 74.12 |
| $g^p$ | 74- I | −51 | | | −40 | | | −47 | |
| | - II | −47 | | | −20 | | | −36 | |
| | -III | −64 | −35 | | −37 | −30 | | −57 | |
| | -IV | −69 | −45 | | −48 | −46 | | −65 | |
| | 75- I | | −26 | −31 | −25 | −23 | 20 | | − 1 |
| | - II | | 17 | 73 | 11 | 27 | 94 | | 50 |
| $u$ | 74- I | − 4 | | | − 2 | | | 0 | |
| | - II | − 4 | | | 4 | | | 8 | |
| | -III | − 7 | − 2 | | 5 | 9 | | 4 | |
| | -IV | −24 | −15 | | − 8 | − 5 | | −12 | |
| | 75- I | | −30 | −13 | −22 | −23 | −16 | | −18 |
| | - II | | −34 | −14 | −25 | −27 | −22 | | −21 |

[a] Table entries are $100[y_\tau(t)-y_\tau]/y_\tau$

where $y_\tau$ = measured state in quarter $\tau$
$y_\tau(t)$ = forecast of $y_\tau$ generated in month t.

[b] Median of forecasts surveyed in *Business Forecasts* (1973–75), Federal Reserve Bank of Richmond.

back strategy to large perturbations indicates the difficulty in efficiently filtering current measurements.

1. *The crowd consolation.* As indicated in Table 2, our hypothetical policy makers were not alone in underestimating the severity of the dislocations of economic activity. Neither the sequence of extravagant inflation rates nor the subsequent jump in unemployment were anticipated with any reasonable lead time by model or judgmental forecasters. In the six quarters prior to 1975, historical monetary policy is not inconsistent with intermittent feedbacking on the effective information sets that were historically available.

2. *The phenomenological phumble.* As might be expected, an untuned model has considerable difficulty in processing stochastic information from unfamiliar sources. Information on the considerable boost in the price of imports and raw materials and dismantling of price controls literally had no place to go in traditional econometric models. (A partial list of structural modifications was suggested in the timely paper by James Pierce and Jared Enzler.) The procedure of simply adjusting structural equation intercepts is not likely to provide efficient filters of perturbations from unexpected sources.

3. *Add factor strategies.* Traditionally, add factor adjustments to model intercepts are intended as corrections to the conditional mean projections of future stochastic disturbances to reflect recent information from model or judgmental data banks. This mean correction procedure is consistent with a certainty equivalence strategy for quadratic loss functions but is inappropriate for asymmetric loss functions such as (1).

It can be shown that an approximation to optimal instrument settings for a simple asymmetric loss function can be obtained by the conventional deterministic solution after the intercepts of the underlying model are adjusted by "add factors" that are proportional to the standard deviations of the model disturbances. Normally,

TABLE 3—REDUCTION OF ONE PERIOD FORECAST ERRORS BY FILTERING CONTEMPORANEOUS INDICATOR ERRORS

| State Variable | Relative Forecast Error Range[a] | Variance Reduction (%) 14 Indicators | Variance Reduction (%) $M_1$ Components |
|---|---|---|---|
| Unemployment rate | 18.9 | 90 | 13 |
| Price rate of change | 43.2 | 58 | 22 |
| Output | 2.7 | 84 | 22 |
| Consumption[b] | 0.8 | 42 | 4 |
| Producers durable investment[b] | 4.1 | 61 | 8 |
| Inventory stock[b] | 1.5 | 52 | 22 |

*Reduction (%) of One-Period Inflation Forecast Error*

| *Filter* | 73–III | 74–I | 74–III | 75–I |
|---|---|---|---|---|
| 14 Indicators | 92 | 15 | 61 | 12 |

[a] Twice standard deviation of forecast error relative to 73–III level.
[b] Constant (1958) dollars.

this strategy adjustment is apt to be of small consequence, but in a period such as 1974 the posterior variance of selected disturbances should be set significantly higher than the estimated variances due to the wider dispersion of prior uncertainty.

4. *A case for monthly filtering.* Since relatively firm estimates of performance in a given quarter are not available until about the middle of the subsequent quarter, in practice there is considerable emphasis on diagnostic examinations of monthly and even weekly entrails to give the planners more elbowroom in the design of the feedback strategy. Indeed, one justification for the prominant role of "judgmental" forecasts in staff briefings is their consideration of weekly and monthly series outside the quarterly model data bank.

An outside indication of the return to massaging intermediate indicators is illustrated in Table 3. The procedure was to determine the potential reduction in the variance of the one-period forecast errors of six candidate state variables, say $e_y$, if one had been able to accurately measure the contemporaneous (quarterly) forecast errors in certain intermediate indicators, say $e_z$. (The significant problem of measurement error is ignored here but see William Conrad.) Since the quarterly model equations are coupled, the forecast error of even a single indicator is potentially a linear combination of all 69 stochastic disturbances in the *MPS* model.

A simple filter was used to "blow up" the measured forecast errors of the indicators $e_z$ into estimates of the disturbances of the structural equations $\hat{\epsilon}$, which were then weighted to form estimates of the forecast errors of the "unobserved" states $\hat{e}_y$.

$$(2) \qquad \hat{e}_y = G_{y\epsilon}\Sigma G'_{z\epsilon}\left[G_{z\epsilon}\Sigma G'_{z\epsilon}\right]^{-1}e_z, \qquad \Sigma = E\{\epsilon\epsilon'\},$$

where $G_{z\epsilon}$ and $G_{y\epsilon}$ are the reduced form multipliers of the disturbances for the indicator and state variables, respectivel.y The variance matrix $\Sigma$ was constructed from a sample of model disturbances for the nine years preceding the feedback interval.

Two sets of indicators were used. The first consisted of 14 variables that can be measured monthly ranging from personal income, manhours and the value of common stock prices to interest rates, components of $M_2$, commercial loans and free reserves. The alternative indicator set consisted only of currency and demand deposits. As indicated in Table 3, $M_1$ alone is not a particularly good indicator of unexpected movements in potential arguments of a policy loss function. It pays to "look at everything" in the context of a well-specified model structure. (See discussion of filtering and references in Kalchbrenner and Tinsley and John H. Kareken, Thomas Muench, and Neil Wallace).

Since variance reduction is only an indicator of average improvement, the bottom of Table 3 indicates how filtering on

the 14 indicator set might have improved the forecasts of price inflation over the feedback period.

We also examined feedback corrections based only on measurement of the forecast errors of the indicator sets. Preliminary results suggest (a) linear approximations of feedback correction rules (i.e., without iterative improvement) can be misleading in the context of a nonlinear model, (b) the effective importance of any given indicator depends on the specification of the loss function, and (c) for the loss function (1), the variance of feedback corrections was dominated by measurements of the forecast errors in free reserves, the bill rate, and the components of $M_2$.

### III. Future Directions

Control theory is neither an autonomic black box nor a miracle cure. Although the returns to filtering are promising, the limited performance of simple feedbacking with historical information sets is a sobering incentive to more creative exploitation of theoretical control techniques.

This paper reports on one aspect of a sequence of control projects in progress at the Federal Reserve. Projects at the Board include (a) the sensitivity of strategies to specification of the policy loss function, (b) modification of the quarterly model to accommodate "rational" expectations, (c) linking of quarterly, monthly and weekly models to efficiently filter current measurements and extract more frequent control settings, and (c) the construction and use of policy multiplier sequences as stochastic processes. (See discussion and references in Kalchbrenner and Tinsley.)

## REFERENCES

**W. E. Conrad,** "Recognition and Response Lags: An Optimal Response to Imperfect Information," presented at Econometric Society winter meetings, Dec. 1975.

**J. H. Kalchbrenner and P. A. Tinsley,** "On the Use of Optimal Control in the Design of Monetary Policy," presented at Helsinki Conference on The Monetary Mechanism in Open Economies, Aug. 1975.

**J. Kareken, T. Muench and N. Wallace,** "Open Market Strategy: The Use of Information Variables," *Amer. Econ. Rev.,* Mar. 1973, *63,* 156–72.

**J. Pierce and J. Enzler,** "The Effects of External Inflationary Shocks," *Brookings Papers,* 1974, 13–54.

# DISCUSSION

HENRY C. WALLICH, Board of Governors of the Federal Reserve System: The use of optimal control techniques in planning for economic stabilization is approaching the policy stage. At the present time, as the papers before us show, the principal application of these techniques has been the examination of models and of past policies. Its use for effective policy advice still seems some distance away. But initial efforts to build an optimal control approach into Federal Reserve policymaking are underway. I believe that there is potential for progress at both the technical and the policy levels. It is important, therefore, for the producers and the potential users of this technique to become better acquainted. Model builders and policy makers must explore one another's needs and capabilities.

Policy makers, I believe, regard their role as somewhat more modest than that with which the terminology of the loss function sometimes endows them. The Federal Reserve, to be specific, is responsible for only one phase of the nation's economic policy—the handling of monetary policy. The overall objectives, moreover, are given by the Employment Act. Most of the economic policies that influence the rate of growth, employment, and the degree of price stability, are handled elsewhere in the government. Particularly when several objectives are involved, which obviously cannot all be attained with one instrument, it seems somewhat presumptuous to state one's preferences in the form of "targets."

The monetary policy official naturally has ideas also about desirable fiscal policy, and about many other policies that influence economic development. Only in the very short run can he make fairly firm assumptions as to what these policies will be. For the longer run, a not unreasonable attitude for him may be to think of monetary policy as helping to create the environment in which other public policies, as well as decisions made in the private sector, will become effective. Any given monetary policy may be consistent with alternative combinations of growth, unemployment, and inflation. The monetary policymaker will adjust his action to what he sees happening in these other spheres. But he should not overestimate his ability to influence the outcome.

The time horizon over which target values are to be set likewise presents difficult problems. One may believe that a lower rate of economic expansion in the immediate future will lead to more sustainable growth and lower ultimate unemployment and inflation than would a more aggressive policy. But unless such preferences are built into a loss function, and a long time horizon is allowed for, rather extreme policy proposals may follow from optimal control techniques applied to econometric models with long lag structures, as some of the papers at this meeting indicate.

The policy maker may also be troubled by an appearance of misplaced concreteness. He may be accustomed to thinking in directional terms—up or down—or in terms of rates of change—faster or slower. He may want to reserve judgment as to precise targets for unemployment and inflation until the economy is a little closer to what he might consider optimal. And if in addition he were asked whether he has a quadratic loss function, i.e., whether he is indifferent to an equal degree of over- and undershooting of his targets, he might be tempted to think the whole thing a spherical nuisance, i.e., a nuisance from every angle.

Monetary policy in the United States, moreover, is made by a group, the Federal Open Market Committee. The loss functions of all its members are unlikely to be identical. Perhaps one could think of a consensus loss function, or at least of one that would be Pareto optimal as between two disagreeing groups. But as a practical matter, it is probably easier for such a group to arrive at agreement on something on which they are compelled to take immediate action such as bank reserves, or the money supply, or the Federal funds rate, than about desirable conditions in the economy over which they have no immediate control. This leads me to my next topic, the role of intermediate targets.

*Intermediate Targets.* In an optimal control framework, it is argued, there is no need for

intermediate targets. It is the ultimate goals that go into the loss function. The pursuit of these ultimate targets by means of intermediate targets such as money supply or interest rates, it is argued, is in theory suboptimal. Their function is to serve as information variables from which insights into developments in the real sector can be gathered. The central bank should look at them as it should look at other readily observable data—everything should be looked at as a source of information about real developments that are not directly or not frequently observable.

This line of argument is in conflict with some of the main developments in monetary policy making in the United States during the last 10 or 15 years. There was a time when the Federal Reserve indeed did "look at everything." It was not a technique lending itself to much precision, and the development of intermediate targets was generally regarded as a step forward.

The advantages of intermediate targets have often been described as those of better observability and better control. I regard particularly the latter as significant. Moving directly from money market conditions or bank reserves, which the central bank controls, to the ultimate targets of growth, employment and price stability conveys very little of a sense of the quantitative impact of monetary policy action. At best, one can have a sense of the direction of policy, and even there, as we have learned, mistakes are possible if the central bank does not distinguish between endogenous and exogenous movements in its policy variables. This uncertainty is enhanced by long lags of policy action. By focusing on the money supply or on interest rates, a better intuitive sense of the thrust of policy is likely to be achieved. The sense of having some protection against the extreme kind of error that might manifest itself by extreme values of these variables, can be helpful.

Policy makers may be reluctant to surrender this means of obtaining some direct feel for the economic meaning of their actions and to replace it by a system that tells them to move some variable like unborrowed reserves, or the Federal funds rate, perhaps drastically, in order to achieve some particular results in the real sector. It might come close to flying with an automatic pilot. Manual control, I believe, would instill greater confidence.

Concern over possibly extreme values of policy variables is likely to be another obstacle to the greater use of optimal control that will have to be dealt with. The experience of the postwar period seems to show that extreme settings of policy variables, even for relatively short periods, can be destabilizing. Frequent variation in instrument settings likewise may add to instability. Given the lags and the uncertainties, moderate instrument settings and a degree of steadiness seems preferable most of the time, quite aside from possible side effects of wide policy gyrations on the functioning of financial markets. Policy makers who do have such preferences for moderation and steadiness can, of course, put them into their loss function, as the paper by Kalchbrenner and Tinsley does. But that is only partial protection, unless the penalties assigned to wide deviations in the paths of instrument variables are very high and hence perhaps unreasonable. Moreover, one may remain suspicious about the properties of a feedback process that needs to be disciplined in this artificial way.

*Models.* The policy evaluation—or advice—derived from an optimal control system presumably is no better than the model through which the feedback flows. Policy makers are likely to be interested in several aspects.

First, while models have reached a certain degree of proficiency in short-run predictions under ordinary circumstances, and in that sense agree with each other, there nevertheless seem to be important differences. Policy multipliers seem to vary importantly among models. Ever within particular models, these multipliers seem to be sensitive to small changes in the specification of particular equations, or to the choice among alternative equations of seemingly equal theoretical and empirical plausibility. Chow's paper suggests a minimax strategy, choosing among the models on the basis of which minimizes the worst case. That procedure would provide some insurance, but otherwise seems to adopt

a rather pessimistic slant. Alternative procedures might be to examine models for robustness of their policy advice under varying assumptions, or perhaps to look for a policy that is robust with respect to switches among models. Clearly one of the precautions policy makers would want to apply is to use a variety of models. But when there are significant differences, it is not easy to work up much confidence.

Second, there is likely to be concern about the possibility of bias in particular models or specifications of loss functions. Ando and Palash point out that in a quadratic loss function, if the target values for unemployment and inflation are set low, such as at zero, the unemployment variable will obtain an unintendedly high weight. My own concern is that on the contrary an inflation bias may enter the process, via the structure and the typical use made of most models. They seem to underestimate inflation because of a questionable process of forming price expectations, which relies on distributed lags of past experience instead of on rational expectations based on observed government policy. The paper by Kalchbrenner and Tinsley makes reference to the severe underestimation of inflation.

Additionally, an inflation bias may appear if model simulations are kept too short, since price effects typically lag volume effects. Long model simulations into the future are not popular, owing to the difficulty of estimating exogenous variables and perhaps also because of the longer run instability of some models. Nevertheless, by limiting a simulation or forecast to only a few quarters ahead, the long-run price effects may be cut off. These effects then will carry less weight in the optimal control simulation while volume effects, which occur with less of a lag, dominate.

Third, still another question about the performance of present-day models relates to their ability to deal with severe exogenous shocks such as the devaluation of the dollar, or the rise in oil prices. Such effects, as Kalchbrenner and Tinsley put it, simply "had no place to go in traditional econometric models." Now that the high rates of unemployment and inflation that resulted have become part of the data, users of the models' output are likely to be concerned about possible distortions from these outliers.

*Uncertainty.* Policy makers may be concerned about the kind of advice they are likely to get when the outlook appears more uncertain than usual. At such times, anyone with a firm opinion is likely to carry disproportionate weight, but in the case of advice from a model that is part of the risk to be guarded against. The natural tendency of policymakers, under such conditions, will be in the direction of greater conservatism, i.e., to do "less" than they otherwise would. The meaning of "less" may not be the same for everyone, although technically it would seem to imply that policy action should then be so designed as not to add to the variance of the loss. In practice, it may just mean to keep doing whatever was being done before.

Among technicians, views do not seem to be unified concerning the implication of varying uncertainty. A well-known theorem by William Brainard states that, under specified conditions, uncertainty reduces the scale of action. Kalchbrenner and Tinsley, in an earlier paper, seem to be of the same opinion. I have heard others quoted to the effect that uncertainty probably but not necessarily implies greater conservatism. In any event, the users of optimal control probably would like to know whether they run the risk, under particularly ticklish conditions, of being confronted with extreme advice from this source as they often are from other sources as well.

*The Outcome of the Policy Simulations.* Ando and Palash and Kalchbrenner and Tinsley have very properly indicated to what extent their findings point to alternative *ex ante* policies that in the light of contemporaneous information could have been adopted to produce better results. *Ex post* simulations, employing information that policymakers did not have at the time, may provide valuable lessons for the future but do not constitute a valid criticism of past policies. If I understand the two pairs of authors correctly, they both claim that, with the benefit of hindsight, policy could have been significantly improved. Kalchbrenner and Tinsley also seem to find that, on an *ex ante* basis, optimal control would not have done better. Feedbacking, in

other words, is not enough to produce better policies, according to their findings. I might add that in a set of papers by Hyman-Shapiro and Hirsch that will be discussed this afternoon, evaluating recent policies with the aid of eight alternative models in an optimal control framework, the conclusion is reached that even with the benefit of hindsight the inflation and recession of the last few years could not have been avoided, although policy could have been improved upon.

GARY FROMM, National Bureau of Economic Research: The papers presented at this and related sessions on optimal control are among the most important being given at the 1975 annual meetings. While none of them is pathbreaking, each contributes to a growing literature in a relatively new field in economics, one that should expand in significance as knowledge and techniques develop over a span of years. At this stage it is easy to find flaws in the methodology and its application for the design of optimal economic policies. But, especially on the eve of the 1976 Bicentennial and Adam Smith's *Wealth of Nations*, it should be recognized that perfection in theory or practice is not to be attained at the outset but only after substantial development and experience.

To some economists, especially policy makers, the thought of using formal models and control theory for the design of optimal stabilization and growth policies seems, on its face, as nonsensical. Yet these same skeptics speak approvingly of automatic stabilizers and formula flexibility, which are highly related to types of strategies that would result from applying control theory methodologies. It should be recognized that at one limit, control theory produces simple strategies like Milton Freidman's prescription of constant money supply growth. But, even if simple rules only are desired, it seems likely that somewhat less rigid control mechanisms could yield improved economic performance. For instance, a slightly less restrictive economic strategy would permit seasonally adjusted money growth rates to vary systematically within prescribed bands inversely to a leading indicator index.

However, even with application of more complex optimal control strategies, complete stability and growth objectives are likely not to be achievable because of lack of knowledge of the structure of an evolving economic system, imperfect and missing data, modeling deficiencies of misspecifications, incompleteness, and aggregation, and stochastic shocks. But, I am certain that the results reported at this session and those of comparable papers are valid—use of control theory techniques can reduce output fluctuations and inflation while raising the realization of economic potential.

With that introduction let me now turn to the present papers. Rather than picking on details of each, I will concentrate my remarks on issues common to all. One of the key aspects is the choice of an objective function. Often, as is the case in several papers here, a quadratic loss function is chosen such that deviations from targets are penalized at nonlinear rates. Such a loss function may, perhaps, correspond to actual preferences. However, analysts frequently choose targets arbitrarily and it is not realized that nonoptimal solutions may be derived as a consequence. This can arise for two reasons. First, with a multiple argument objective function, if all targets are not achievable, points within the possibility frontier are as likely to be selected as solutions as points on the frontier. Second, unless the targets are on the frontier at the bliss point, target achievement does not lead to maximization of social utility (or, the converse, minimization of disutility) nor to optimum policy. That is, a target loss function is appropriate only if targets are at the bliss point. Determination of the latter requires knowledge of both the structural constraints of the system (the frontier) and the positive utility function. But, if the positive utility function is known (or assumed), then it no longer is necessary to ascertain the bliss point in advance, and direct solution for optimal policy can be undertaken without specifying any targets or goals.

Clearly, the economics profession must do considerably more work on the determination of social utility functions before strong reliance can be placed on optimal control policy pre-

scriptions. Again, despite the fears of some skeptics, this is not as hopeless as it seems at first blush. Both direct survey and indirect revealed preference methods already have shown some promise in ascertaining social trade-offs or preferences for output, inflation, income transfers, and related variables. It should be remembered, too, that even if social preference functions are not explicitly identified, they are implicit in economic policy choices made by the President and the Congress. The key question is can we do better by consideration of explicit trade-off functions than by leaving them hidden and implicit. The answer, I believe is clear. As in other areas of decision analysis, formalization of implicit assumptions and structural characteristics (the trade-off functions and constraints) should lead to more informed and intelligent choices which are more consistent with welfare maximization than those which would occur without such explicit identification.

In specifying or estimating these welfare or tradeoff functions, analysts should beware of strong simplifying assumptions such as ignoring time dimensions and preferences, complementarities in preferences among arguments, trade-offs in amounts and variations in arguments, and, in general, the impact of uncertainty and stochastic influences. Too often, for the sake of expediency in the ease or cost of computations, extremely simplified functions have been assumed. There is need to perform sensitivity analyses with different forms, arguments, and parameters of such functions because policies and their ranking can change dramatically as these are altered. Fromm and Taubman found, for instance, that choices between policies shifted as the elasticity of substitution in a *CES* utility function was varied within plausible limits.

The same violence of simplications has been true with regard to structural constraints. If the world truly were linear and recurcive, optimal control solutions could easily be calculated. But most models of the economy reveal significant nonlinearities and, in selected blocks, strong simultaneity. My own experience with the Brookings and *FED-MIT* models has shown that linearizing such systems, even on a piecewise basis, drastically alters structural characteristics and multipliers on policy instruments. Therefore, extreme caution is advised in interpreting optimal control results that depend heavily on linear approximations.

The handling of uncertainty is another problem area. Again, based on simulations with the Brookings model and employing a utility function that incorporates risk preferences, it would appear that strong simplifications, such as certainty equivalence, are dangerous and lead to nonoptimal policy prescriptions. One clear need in this area is to define some terminology. For instance, we need definitions that differentiate between stochastic terms in estimation and model solution within sample periods, beyond sample periods, and for residuals of *ex ante* and *ex post* predictions with and without introduction of prior information such as constant and parameter adjustments or add factors.

In the short time allotted, it has been possible only to touch on a few general issues and not grapple with details and some of the admittedly strange results of applying control techniques in the Ando and Palash and Kalchbrenner and Tinsley papers. But these authors and Gregory Chow, are to be congratulated for fine efforts which should lay a basis for much future work in this important field.

POLITICAL ECONOMY: SOME USES OF THE EXIT-VOICE APPROACH

# Individual Mobility and Union Voice in the Labor Market

*By* R. B. Freeman*

Standard economic analysis of the impact of trade unions on the labor market is straightforward: unions are monopolistic organizations that raise wages and create inefficiency in resource allocation. Industrial relations experts tell a more complex story, stressing the diverse effect of unions on work rules, managerial decision making, and "virtually every aspect" of the activity of enterprises. To what extent do the non-wage effects of collective bargaining make the monopoly model incomplete or misleading? Do unions perform economically significant functions beyond altering wage rates?

This paper examines the non-wage effects of trade unions in the context of the "exit-voice" model of the social system developed by Albert O. Hirschman (1970). Unions are treated as the institution of collective voice in the job market; voluntary quits as the expression of exit. Comparisons are made between the free market quit and union voice mechanisms for transmitting worker desires for conditions of employment, compensation packages, and rules of the work place to employers; some empirical implications of the analysis are drawn and preliminary empirical findings described.

* Dept. of Economics, Harvard University. I have benefited from the comments of Albert Hirschman, James Medoff, John Dunlop, Zvi Griliches, and Martin Segal.

## I. Employment Relations and Conditions of Work

What makes the exit-voice framework a potentially fruitful way of looking at trade unions are "the peculiarities of the employment contract . . . which distinguish it from other kinds of contracts." (Herbert Simon, p. 183) In a world of uncertainty, imperfect information, and transactions costs, employment involves an authority relation in which workers sell "labor power" to enterprises for extended periods of time. As Ronald Coase puts it, "the contract is one whereby the factor (labor) for a certain remuneration . . . agrees to obey the directions of an entrepreneur *with certain limits*" (pp. 336–37). Because of on-the-job skills specific to enterprises (a learning curve phenomenon) and the cost of mobility and turnover, there are gains to be had from regular employment, a continuing relation between firms and much of their work force, in which allocative and remunerative decisions are not directly determined by the price mechanism.[1] Because workers care about nonpecuniary conditions of employment and rules of the work place, and because different conditions, rules, and methods of organization have different costs, the labor contract tends to be com-

[1] Because I focus on the situation with regular work relations, the analysis is not directly applicable to the case of craftsmen who change jobs regularly.

plex and multidimensional, involving numerous issues beyond wages, including the "social relations of production." (John R. Hicks, pp. 317–18) Because workers have some control over their own activities and can affect the productivity of others, particularly in "team" settings where monitoring of individuals is expensive, their attitudes or morale are potentially important inputs into the production process.

The complex, multidimensional and continuing nature of the employment relation creates a substantial information problem in the job market (Oliver Williamson, Michael Wachter, and S. Harris). Within firms, prices convey only crude information about preferences and costs. In the external market, workers and employers must appraise diverse conditions, rules, and compensation packages across enterprises. Over time technological changes continually alter the structure of the work place and jobs while changes in real income lead workers to demand new conditions of employment and methods of compensation. Firms or organizations have continual need for information from workers not only as Hirschman stresses because of decline or decay but because of changes in opportunities and efficient modes of behavior.

From the perspective of the exit-voice framework information about conditions and preferences can be provided by: free market "exit," consisting of quits and related behavior and/or by "voice" consisting of the collective bargaining system by which workers elect union leaders to represent them in bargaining. Each mechanism has certain strengths and weaknesses, which must enter any welfare economics calculation. Each is, from the perspective of positive economics, likely to produce different market outcomes in the various aspects of the labor contract.

## II. Quits as an Information System

Worker control of their own effort in work places creates several forms of "exit-behavior": quitting, rejecting a job offer; absenteeism, the partial withdrawal of labor time; reduction of work effort in the form of malingering and slack or in extreme cases "quiet sabotage." Strikes, which involve temporary collective exit, are best viewed as a tool of voice since they involve expression of demands by union or union-like group organizations.

Employers can learn about worker preferences and the causes of discontent from individual quits either inferentially by linking different levels of exit to different or changing characteristics of the employment relation in what amounts to hedonic price type calculations or through direct questioning, "exit interviews," of those who quit or make related decisions.

The amount of information provided by either channel is, unfortunately, likely to be small relative to the costs of processing the data. Inferring reasons for discontent from "abnormally" high quits requires specification of underlying multidimensional characteristics of work places; a reasonably large sample; and sufficient variation in conditions or in changes in conditions to permit the assignment of causality. Inferences about the preferences of workers who quit may, because of "selectivity bias," yield incorrect information about average evaluations or the evaluations of future potential quitters. The wider the variation in tastes among workers due to differences in preferences and position in the work place (especially seniority) and the greater the different modes of expressing exit (the young quit, while older employees choose less drastic actions), the more difficult will the inferential process be. If all persons had the same tastes, introspection by managers or foremen would yield answers at no cost.

Exit interviews run into a different difficulty—that of motivating the worker who leaves to detail work place problems. There is no gain to quitters from providing management with desired information about dissatisfaction, and there are possible losses via bad references—retaliation for conveying bad news. As a result, the leading personnel management text reports that "it is extremely difficult to get reliable information by means of exit interviews" (C. Meyers and P. Pigors, p. 222). The information content in other forms of exit behavior will be even smaller. A worker guilty of absenteeism for reasons of discontent with work relations will hide the reasons for fear of being fired, while malingering or quiet sabotage are by their nature covert activities.

Finally, whatever information the quit mechanism provides about current conditions of work, it provides much less about changes in conditions and, more importantly, about the trade-offs workers are willing to make in wages or other conditions for desired improvements; especially when the changes involve more than marginal adjustments in the work place. Innovations in the labor contract can be appraised only by potentially costly trial and error experimentation: changing conditions or compensation and observing outcomes.

Despite the high noise to signal ratio in labor market quits, the mechanism has desirable properties. Quit behavior depends solely on individual free market "marginal decisions" and does not require collective organizations like unions. Such behavior has well-known efficiency properties. Over time, some employers will extract the signal from the noise in quits or by trial and error find the "optimal" labor contract and come to dominate production or be imitated by competitors.

Certain features of the quit mechanism suggest, however, that there may be some divergence from optimum for extended periods of time. Since workers who leave an enterprise do not benefit from the improvement in conditions that may result, the extent of quitting to convey information may be suboptimal. That is when quits are effective in altering work conditions, the irreversibility of the decision provides too little incentive for workers to leave. Even if the level of *quitting* were optimal, however, the possibility that quitters differ in their evaluation of conditions from other workers (notably older or specifically trained persons who are unlikely to quit and whose position in the firm gives them a different set of interests) makes reliance on quits potentially misleading. Other forms of exit behavior must also be evaluated. In recessions or declining markets, moreover, quits will be ineffective as a mode of workers expressing preferences and obtaining desired changes.

Hirschman's analysis suggests other malfunctioning. Consider the situation in which all employers offer the same work conditions and where workers quit particular enterprises in the hope of attaining better conditions but in fact cannot do so. Each firm has a constant pool of applicants for jobs and a constant quit rate; no information is conveyed about what to change and there are no advantages in profitability. More generally, the same result would occur if the firms had different but similarly undesirable work conditions. Workers disliking bad condition $A$ in firm $A$ would move to firm $B$ where they would find bad condition $B$ and so forth. In this musical chairs equilibrium, quits are ineffective and excessive, at least until "random mutations" by enterprises produce a better wage-benefits-conditions package, so that quits differ substantially among employers. As a result there will be a long

period of adjustment to equilibrium.

From the perspective of workers, the information flow is especially likely to be faulty with respect to aspects of jobs, ranging from treatment by supervisors and coworkers to actual (as opposed to nominal) work responsibilities to evaluations of hazard (W. Kip Viscusi), which cannot be calculated without actually accepting employment. Under the quit system, each worker, would, say, learn about the bad qualities of a job only after a trial period of work and quit. If all workers knew about the job, no one would apply (without larger compensating differentials), but the information each gains is lost.

The question is not, however, whether individual mobility *ultimately* provides enough information to attain something akin to the optimal labor contract in the long run but rather how it does relative to the alternative institutional mechanism of voice, to whose operation I turn next.

## III. Collective Voice by Trade Unionism

The institution of voice in the labor market is trade unionism and collective bargaining. There are several reasons why collective rather than individual activity is necessary for voice to be effective within firms. First, the authority relation makes it difficult for individuals to express discontent due to the danger of being fired—it is clearly easier to retaliate against a single worker than the entire work force. Even with unionization, however, some protection of voice is needed, particularly for activists, as is recognized in the Wagner Act, which makes it an unfair labor practice to fire or otherwise discriminate against persons for trade union activity. Second, the communal nature of work conditions and rules which apply to all workers in the establishment creates a public goods problem of preference relation. Individuals cannot bargain over plant-wide conditions nor trade wages for fringes when there are sizeable set up costs. They will not have the incentive to reveal their preferences when the activity of others may produce the public goods at no cost to them. Elicitation of preferences and determination of acceptable bargaining packages are one of the major tasks of trade unions and critical components in the operation of successful collective bargaining systems. Third, because of the regularity of employment, there is a need to "police" or monitor contracts and thus for a collective agency specializing in information about the contract and in representation of workers. The ubiquity of grievance systems (found in 96 percent of contracts) or grievance and arbitration (93 percent) reflects the importance of contract interpretation in the operation of collective voice (U.S. Department of Labor).

The major advantages of unionization are that it provides: a direct channel of communication between workers and managenment; an alternative mode of expressing discontent than quitting, with consequent reductions in turnover costs and increases in specific training and work conditions; and social relations of production which can mitigate the problems associated with the authority relation in firms.

Union voice proffers very different information about workers' preferences than individual mobility—specific facts about areas of discontent and actual tradeoff possibilities (albeit masked by negotiating strategies). It creates an institutional mechanism for innovation in labor contracts and what may be termed a "new market" for labor contracts.[2] It focuses managerial attention and effort on labor

[2] Recent theoretical work on the existence of contingent claims markets and constrained transactions is thus relevant, surprising as this may be to the theorists, to analysis of the trade union institution (Jerry Green and Eytan Sheshinski).

issues on a regular basis, alerting firms to problems fairly continually rather than by sudden sharp outbursts of discontent. In large enterprises, union voice provides central management with information about local conditions and operations that is likely to differ greatly from that obtained from the organizational hierarchy.

By providing a mode of expressing discontent beyond exiting, direct information about worker desires and certain preferred work conditions that cannot be readily offered by nonunion establishments, union voice can be expected to reduce quit rates, absenteeism and related exit behavior. In the exit-voice model, there is *ceteris paribus,* a reasonably well-defined and economically significant trade-off between the two mechanisms. The reduction in quits will reduce labor turnover and training costs and increase firm-specific investments in human capital and possibly have efficiency gains.[3]

One desirable nonpecuniary condition of work, an industrial jurisprudence system with formal grievance, arbitration and related protection against managerial authority in establishments, may be "produced" only by unionization or some related mechanism. The essence of industrial jurisprudence is the dilution of the power of the foremen and other supervisors, which would presumably be difficult to attain in nonunion establishments where management has the "last word." Desire for such rules and procedures implies that union voice enters as a direct argument in utility functions (suggesting that in the absence of monopoly wage gains, workers would take lower pay for collective bargaining) and by affecting morale and the social relations in the firm, possibly in the production function as well.

On the negative side, the major disadvantage of union voice is that, as the standard monopoly model stresses, it creates monopoly power in the job market and wages above and employment below competitive levels. There are other costs to unionization as well: government, management, and workers invest considerable resources in bargaining which might have been spent on direct production; the productivity of the work force may be reduced by "featherbedding" (one possible exercise of monopoly power which would not show up in wages per se). Some work rules, like closed shops, may be "inequitable" to some workers; and more generally as a "political organization" unions may not reflect the desires of all members in a reasonable or "fair" manner.

Accepting the costs of collective bargaining, the normative question is whether these costs exceed the benefits due to reduction in exit behavior, improved flow of information, and provision of protection against managerial authority. There is, in our analysis, a trade-off: by providing information and a mechanism for potentially complex bargaining among workers and between workers and management that is more efficient than quits, collective bargaining is likely to yield a better "mix" among wages, work conditions, rules of the work place and a reduction in turnover costs and increased firm specific human capital, at the expense of higher price of labor.

## IV. Union Voice and Market Outcomes

The fact that unions are "political" organizations whose activities depend on the preferences of workers as a group has important implications regarding the impact of collective bargaining.[4] Whereas quits reflect the desires of marginal work-

[3] Of course the reduction in quits could be too great with bad consequences for overall labor mobility.

[4] While recognizing the political and organizational nature of unions, in this essay I neglect the effect of these factors on behavior, treating unions solely as an institution that conveys the desires of members. This omission will be rectified in future work.

ers, voice reflects the demands of some average of workers. If as a first approximation the median voter model is applied to union behavior, policy will be set by the median member (who is the marginal voter) with the consequence that greater weight will be placed on the preferences of relatively immobile workers (such as older workers and the specifically trained) than under the quit mechanism.[5] Trade unionism transforms the supply side of the job market by *making median (or some other average) rather than marginal preferences the "determinant" of the labor contract.* In the job market, individuals for whom there is a great gap between (pecuniary and nonpecuniary) wages in an establishment and opportunity wages are especially likely to exercise voice when work conditions change in an undesirable way. This contrasts with Hirschman's product market case in which those with higher consumer surplus are most likely to exit when quality of goods deteriorate.

The switch from a marginal to median (or other average) supply calculus might be expected to create inefficiencies since optimality criteria invariably require marginal first order conditions. Because of the union's role in facilitating information flows and coordinating worker preferences in bargaining, however, the situation is more complex. Bargaining among workers (which would not exist in the absence of unions) and potential worker demands for public or lumpy goods at the work place *could* offset inefficiency losses. Consider, for example, a situation in which management can choose one of two modes of organizing work, which, exclusive of their impact on workers, have equal profitability. Method A greatly reduces the well-being of immobile senior workers, while method B has no effect on them but displeases the mobile young slightly. In a market where information is conveyed by quits, the behavior of the young would lead management to choose method A, despite the loss of consumer "surplus" to older personnel. In a market with collective bargaining, the union might arbitrage the differences in preferences, so that the firm will pick B, with a negotiated redistribution giving the young some compensatory benefit and the old a less onerous loss than under A. This scenario can be expanded, to take account of different frequencies of quitting under A and B with similar results. The greater the difference between the losses under the two modes, the greater is the possibility that a superior bargain could be struck through the voice mechanism. I do not claim that the union will, in fact, arbitrage worker preferences correctly for the behavior of the union will depend on its internal organization, organizational goals, and political power of the various groups which are neglected here. The possibility is, however, there.

A second situation in which average preferences may yield better outcomes occurs when desirable work conditions or fringe benefits involve substantial fixed costs. Marginal evaluation could, as is well-known, lead enterprises to reject provision of employment conditions which would pass a benefit-cost test.

Advantages or disadvantages of union voice aside, by providing a distinct market mechanism for imparting information, aggregating preferences, altering authority relations, and changing marginal to average evaluations, collective bargaining will yield outcomes that differ from those in competitive markets. I conclude by sketching briefly some of the empirical implications of the model and the evidence regarding them.

[5] The median voter model is used as the simplest voting model. With many dimensions to contracts and groups within unions, it is obvious that a more complex game-theoretic model is needed to explain actual events.

As noted earlier, unionization is expected to reduce quits and raise investments in firm specific training. Studies of interindustrial variation in quits have generally found a negative relation (John Pencavel, John M. Burton and J. Parker, Vladimir Stoikov and Robert Raimon), as does my preliminary analysis of the relation between unionization and quits using disaggregate data files, which provide a superior test, through the route of the linkage remains to be determined. At least one potential proxy for "specific training," years of tenure with an employer, is also positively affected by unionization (Freeman, forthcoming, a).

Reversing lines of causality, the exit-voice model suggests that when quitting is nonviable, workers may be more prone to unionization. This is roughly consistent with Alan Blinder's evidence that older workers and those with family responsibility are more likely to be union members; with *U.S.* trade union history, in which unionism has grown rapidly following major recessions or depressions; and with union activities in coal, lumber, and similar locales from which exit is costly. Other factors are obviously also important, however.

Trade unionism can be expected to increase the fringe benefits share of total compensation, both for information considerations (Richard Lester, p. 494) and preference aggregation reasons. Analysis of data from individual establishments finds a significant positive relation (Freeman, forthcoming, b).

As a relatively permanent market institution, preserving information about work conditions that might otherwise be lost through individual quits, trade unions can also be expected to *increase* differentials for bad working conditions relative to the competitive market outcome, a result supported by Viscusi's analysis of compensating differentials for dangerous work.

Through the various nonwage effects outlined here, unions might be expected to alter the overall production process and worker productivity. James Medoff and C. Brown find, in fact, positive trade union productivity effects in cross-industry, cross state regressions, holding capital-labor and worker quality fixed.

Finally the model suggests that by directing attention to work place problems and encouraging expressions of discontent and by keeping dissatisfied workers from quitting, unions may *increase* job dissatisfaction, other factors (wages, conditions) held fixed. My preliminary investigation of job satisfaction in large data tapes finds such a relation.

If the unionization-satisfaction result stands up to more detailed analysis, it has significant implications for understanding the entire voice mechanism. It suggests that voice, of necessity, produces "dissatisfaction" by making individuals especially sensitive to and willing to criticize conditions. To be effective, voice cannot be silent, even when it may produce "the goods."

## REFERENCES


A. **Blinder,** "Who Joins Unions," working pap. no. 36, Industrial Relations Section, Princeton Univ.

C. **Brown and J. Medoff,** "Trade Unions in the Production Process," unpublished pap., 1975.

J. **Burton and J. Parker,** "Interindustry Variations in Voluntary Labor Mobility," *Industrial and Labor Relations Rev.,* Jan. 1969, 199–216.

R. **Coase,** "The Nature of the Firm," Ch. 16 in American Economic Association *Readings in Price Theory,* Vol. VI, Chicago 1952, 331–51.

R. **Freeman,** "The Impact of Unionization on Quits: The Tradeoff of Exit and Voice in the Job Market," forthcoming, a.

———, "The Effect of Trade Unions on Fringe Benefits," forthcoming, b.

**J. Green and E. Sheshinski**, "Competitive Inefficiencies in the Presence of Constrained Transactions," *J. Econ. Theory*, June 1975, *10*, 343–57.

**J. Hicks**, *Theory of Wages*, New York 1973.

**A. O. Hirschman**, *Exit, Voice and Loyalty*, Cambridge 1970.

**R. Lester**, "Benefits as a Preferred Form of Compensation," *Southern Econ. J.*, Apr. 1967, *33*, no. 4, 488–95.

**H. G. Lewis**, "Competitive and Monopoly Unionism," in *The Public Stake in Union Power*, Charlottesville 1959.

**C. Meyers and P. Pigors**, *Personnel Administration*, New York 1973.

**J. Pencavel**, *An Analysis of the Quit Rate in American Manufacturing Industry*, Princeton 1970.

**H. Simon**, "A Formal Theory of the Employment Relation," in *Models of Man*, New York 1957, 183–95.

**V. Stoikov and R. L. Raimon**, "Determinants of Differences in the Quit Rates Among Industries," *Amer. Econ. Rev.*, Dec. 1968, *68*, 1283–98.

**U.S. Dept. of Labor**, *Characteristics of Agreements Covering 20,000 Workers or More*, bull. 17.29, Washington 1972.

**W. K. Viscusi**, "Employment Hazards," dissertation in progress, Harvard Univ.

**V. Vroom and E. Deci**, *Management and Motivation*, Penguin 1974.

**O. Williamson, M. Wachter and S. Harris**, "Understanding the Employment Relation: Analysis of Idiosyncratic Exchange," *Bell J.*, Spring 1975.

# The Economics of Internal Organization: Exit and Voice in Relation to Markets and Hierarchies

*By* OLIVER E. WILLIAMSON*

The economics of internal organization is concerned with two closely related issues. As contrasted with received micro theory, which is preoccupied with the operation of markets (whence the firm exists mainly as an analytical convenience), the economics of internal organization emphasizes that market organization and hierarchical (internal) organization are alternative—sometimes substitute but also complementary—means of executing transactions. The object is to assign transactions to markets and hierarchies so as to achieve a most preferred (usually least cost) result, where this is judged principally in terms of transaction costs. Whereas conventional micro theory is relatively unconcerned with assessing the merits of executing transactions by one mode rather than another (partly because prevailing modes are commonly held to be optimal), the economics of internal organization regards the assignment of transactions to markets and hierarchies as intrinsically interesting.

The second issue of importance to the study of the economics of internal organization concerns organization form. Just as market structure matters in assessing the efficacy of price mediated transactions in the market place, so internal structure matters in assessing the properties of administrative organization. The hierarchical decomposition of tasks and the types of internal control processes that are employed are both germane to an assessment of internal organization.

This paper sets out some of the leading attributes of Albert Hirschman's *Exit, Voice and Loyalty* (*EVL*) (1970) and my *Markets and Hierarchies* (*MH*) (1975). As will be evident, both sometimes complement received micro theory where conventional theory is silent or ambivalent. In other respects, however, *EVL* and *MH* offer substitute ways of interpreting or dealing with issues that fall within the range of received micro theory—or straightforward applications thereof.[1] *EVL* and *MH* also bear complementary and substitute relations to each other.[2]

## I. Leading Attributes

Both *EVL* and *MH* approach the phenomena of interest in a "rational spirit." Rationality is construed somewhat more broadly, however, than is characteristic in

* Research on this paper was supported by a grant from the National Science Foundation (NSF SOC 72-05550 A02). Helpful comments by Albert Hirschman on an earlier version of the paper are gratefully acknowledged.

[1] Although stilted, received micro theory will be taken to be the approach to the study of economics set out in any of the better intermediate micro theory textbooks.

[2] There is a presumption that the reader is already familiar with *EVL* and *MH*, either as a result of having read the books or one or more papers that relate thereto. A series of papers that develop *EVL*, including a survey by Hirschman, were published in *Social Science Information* during 1974. Also see R. R. Nelson and M. Krashinskey (1973). Papers dealing with *MH* include Oliver Williamson (1974a); Williamson, Michael Wachter and J. Harris (1975); and Williamson (1976).

economics: although net benefit considerations govern, extra-economic and systems considerations, both of which stretch the net benefit calculus beyond its normal range, also appear.

### *A. Analytical Novelty*

Given that a good or service is to be provided, *EVL* is concerned with the quality characteristics of the resulting product—with special attention to the question of whether potential spillover benefits are conferred on the less informed part of the population by the more informed subset. *EVL* proposes that the standard economic model, in which "exit" is the principal mechanism by which adaptations are accomplished, be augmented to make allowance for the mechanism of politics; namely, "voice." Exit adaptations can and usually are taken by parties acting independently. For voice to be effective, however, generally requires collective action.[3]

Rather than focus on goods or services as such, *MH* is concerned with the manner in which the implied set of transactions needed to accomplish any given productive outcome are organized. Contractual efficiency requires that transactions be organized in such a way as to economize on the bounded rationality attributes of human decision makers and to design interfaces between successive stages of production that promote harmony and attenuate opportunism. Both market and internal contracting modes are examined. Market modes include complex contingent claims contracting, incomplete long-term contracts, and recurrent spot contracts. Internal modes include peer group organization, simple hierarchies, and compound hierarchies of various kinds. As contrasted with the standard economic model of received micro theory, in which the firm is regarded as a production function to which a profit maximization objective has been assigned, the firm in *MH* terms is a device for economizing on transaction costs. For such economizing purposes, the details of internal organization (in factoring problems hierarchically and in internal control respects) matter.

### *B. Paradigm Problems*

The Nigerian railway case may be regarded as the paradigmatic example in *EVL*. Hirschman puts it as follows:

> The presence of a ready alternative to rail transport makes it less, rather than more, likely that the weaknesses of the railways will be fought rather than indulged. . . . This may be the reason public enterprise, not only in Nigeria but in many other countries, has strangely been at its weakest in sectors such as transportation and education where it is subjected to competition. [p. 44]

The suggested remedy involves blocking or inhibiting exit, thereby to mobilize voice and achieve a collectively superior outcome. Variations on the argument appear in conjunction with the treatment of public schools in the United States and of oligopoly.

The paradigmatic example in *MH* is the insurance problem. As is well-known, insurance suffers from disabilities of two kinds: adverse selection and moral hazard. It is less widely appreciated, however, that both are attributable to the same set of causes. Thus adverse selection difficulties appear because of (1) the impossibility or high cost of making *ex ante* discriminating judgments between applicants whose objective risks differ, which is a manifestation of bounded rationality, and because (2) poor risks cannot be relied upon to disclose their risk characteristics accurately, which is a reflection of oppor-

[3] *EVL* also points out that the study of politics (voice) can benefit by paying greater attention to the use of exit—as an alternative to or used in conjunction with voice. This aspect of the argument is neglected in this paper.

tunism. Similarly, moral hazard problems appear because of (1) the high cost of assessing *ex post* insurance claims in circumstances where risks and decisions are confounded (bounded rationality) and because (2) reports by insureds of negligent behavior cannot be taken at face value (opportunism). Conventional insurance is less complete for these reasons than it otherwise would be and nonmarket or market-assisted insurance alternatives are correspondingly more important.

By simple reinterpretation of parameters, in which productivity attributes of workers are substituted for risk attributes of insureds, the employment relation can be shown to be of the same form as the insurance problem; problems akin to both adverse selection and moral hazard arise and for identical reasons. The vertical integration problem, in turn, is a variant on the employment relation; and strong parallels exist or can be established in studying conglomerate organization as well. Thus labor markets, intermediate product markets, and capital markets can each be interpreted in a consistent framework. Hierarchical modes of organization supplant market modes in each instance in the degree to which they serve to economize on bounded rationality and attenuate the exercise of opportunism.

### *C. Antecedent Literature*

The antecedent tradition on which *EVL* relies is a mixture of politics and economics, the former of which is rather diffuse. Within economics, *EVL* relies in part on the literature that has been concerned with organizational slack (Richard M. Cyert and James G. March; Harvey Leibenstein) and partly on the market failure literature, of which Thomas Schelling's subsequent work on the "ecology of micromotives" is an example. This last is especially germane, in that the suboptimization outcomes which sometimes result when micromotives operate in an unrestrained way is the major concern of *EVL*.

The antecedent literatures on which *MH* mainly relies are (1) the transaction cost literature of the 1930's (John R. Commons; Ronald Coase), (2) the more recent market failure (Kenneth Arrow 1969) and contingent claims contracting literatures (Arrow 1974, Ch. 2), (3) aspects of organization theory (Herbert A. Simon, I. Goffman) and (4) the importance to the study of economics of factor nonhomogeneity (Friedrich A. Hayek, Peter B. Doeringer and Michael J. Piore). An attempt is made to identify a core set of factors which, directly or indirectly, underlie the transactional frictions with which each of these literatures is separately concerned. An "organizational failures framework" is then proposed, which framework is applied to the study of transactions across markets and within firms in a variety of contexts. The pairing of bounded rationality with uncertainty/complexity on the one hand and of opportunism with small numbers bargaining conditions on the other are key features of the framework.

### *D. Extraeconomic Dimensions*

Both *EVL* and *MH* introduce extraeconomic factors that do not ordinarily appear in conventional analysis. This does not imply that efficiency considerations are suppressed; rather a broad definition of efficiency is favored in which it is acknowledged that extraeconomic factors matter and by noting that alternative modes of organization affect these differently.[4]

A central concern of EVL is that market (exit) alternatives can *"atrophy the development of the art of voice"* (p. 43,

[4] In neither case, however, are the extraeconomic factors that are introduced well-integrated into the analysis.

emphasis in original), with debilitating system consequences. This challenges the usual economic presumption that more degrees of freedom are always to be preferred to fewer. The Hirschman view is shared by R. M. Titmuss, who argues that an all-donor system of collecting blood has superior properties to a mixed voluntary-commercial effort (Titmuss). The circumstances under which such arguments can be expected to hold, however, have yet to be delimited. The relevant public policy subset to which *EVL* applies is thus not well defined.

*EVL* also introduces the notion of loyalty and is concerned with its consequences. When is this functional, in that it facilitates recovery from adversity, and what are its limits? Finally, although *EVL* originally interpreted the exercise of voice as a costly undertaking, Hirschman now acknowledges that—for some individuals, for some purposes, over some range—voicing itself is valued. Evidently dissent has cathartic effects and/or the prospect of influencing outcomes is affirmatively regarded. "Public interest" activities for which the voicer can expect to receive acclaim attract voicing energies.

*MH* introduces the notion of atmosphere and argues that supplying a satisfying exchange relation is part of the economic problem, broadly construed. Reference to atmosphere is intended to make allowance for attitudinal interactions, and the systems consequences that are associated therewith. Market exchange tends predominantly to encourage calculative relations of a carefully metered sort and works well when there are few unsettled obligations that carry over from one transaction to the next. To process all transactions as though they were of this kind, however, is unwarranted. In particular to mediate transactions which are linked by attitudinal nonseparabilities as though they were independent (hence can be settled in a transaction-specific, *quid pro quo* fashion) can lead to inferior contracting modes in which anonymity is overstressed and system losses are incurred. Contracting for the services of human assets is the leading instance where the influence of current metering on future performance needs to be assessed with care. Considerations of internal due process arise in this connection. Collective organization, in both peer group and employment relation respects, arises partly for this reason.

## II. Some Similarities

### *A. New World Views*

The authors of *EVL* and *MH* plainly believe that they are onto a new way of organizing a wide set of social science phenomena. Hirschman puts it as follows: "I had come upon a manner of analyzing certain economic processes which promised to illuminate a wide range of social, political, and indeed moral phenomena" (p. vii). Among the applications which he notes are "issues as diverse as competition and the two-party system, divorce and the American character, black power and the failure of 'unhappy' top officials to resign over Vietnam" (p. vii). Likewise Williamson observes that "I was advised by a number of scholars to whom I distributed an earlier version of the manuscript that a distinctive worldview was being offered and that this ought to be emphasized" (p. xii), and goes on in his concluding chapter to contrast the proposed approach with a series of alternative paradigms—including, in addition to received micro theory, the structure-conduct-performance paradigm and the property rights approach.

That the *EVL* and *MH* approaches generalize is evident from subsequent work that has been done in each. The two other papers given in this session demonstrate that *EVL* helps to illuminate aspects of

both labor and urban economics. For an application of *MH* to the study of regulation see Williamson (1976).

### *B. Voice*

Both *EVL* and *MH* feature the use of market alternatives to autonomous contracting in which voice is prominent. But voice can take a variety of forms, and the forms characteristic of *EVL* and *MH* are rather different. Thus voice, as it is used in *EVL*, is mainly a protest activity in which consumers are appealing to suppliers for better quality products. Voice here takes place across a market. In *MH*, by contrast, voice takes place among the members of an internal organization and occurs in a continuing rather than in a sporadic way. Adaptations of a sequential-decision making kind, rather than of a prespecified contractual kind, are thereby facilitated.

Internal organization can be regarded, for some purposes, as an effort to attenuate costly and repeated bargaining between autonomous parties by joining the two parties in a common ownership unit, thereby to achieve a more nearly joint profit maximizing outcome. The resulting internal dialogue is apt to be rather different than that which occurs between a supplier of a good or service who is confronted by an articulate group of dissident customers who seek a change in policy which favors their objectives. Only if the dissidents were to integrate backward or take over the supplier would the parallels between *EVL* and *MH* be reasonably complete. This is not, however, in the spirit of *EVL*.

### *C. Understatement of Market Powers*

Both *EVL* and *MH* are of special relevance in circumstances where markets either do not or prospectively will not work well. The emphasis, accordingly, is on market frictions and on devising non-market or market assisted adaptations thereto. This is altogether useful, but the remarkable properties of markets for bringing large numbers of buyers and sellers into adjustment are somewhat neglected in the process. An undervaluation of the powers of markets may be an unintended byproduct of both.

## III. Some Differences

The differences that are treated here are often not very great. Sometimes they serve to complete or elaborate remarks already made in the sections above.

### *A. Markets*

#### 1. *Oligopoly*[5]

*EVL* takes the position that firms in oligopolistic industries are able and often do offer their customers shoddy merchandise without either losing trade or suffering the complaints of voice. Disgruntled, but silent and nonknowledgeable, customers move from one oligopolist to another under the "perpetual illusion" that things will get better only to discover, to their dismay, that they do not. The apparent exit choices that oligopoly affords are accordingly characterized as wasteful and diversionary. Full monopoly is said to be better since, absent the illusion of exit, voice would be mobilized and superior results would obtain.

There are two problems with the argument. For one thing, consolidated (monopoly) ownership gives no assurance that the diversionary range of choice to which Hirschman objects will be altered. If the monopolist offers multiple brands—which in the automobile industry to which Hirschman refers (pp. 27–28) is to be expected—voice-blunting consequences will presumably continue. Except, therefore, as product differentiation is impossible (as

[5] The discussion in this subsection follows Williamson (1974a, 1974b).

in homogeneous product industries), is prohibited, or users are somehow locked into whatever brand they are initially assigned, the voicing response on which Hirschman relies may fail to materialize. Additionally, Hirschman appears to undervalue the benefits of oligopolistic competition. *MH* offers a rather different view, where the problem of oligopolistic collusion is posed in contracting terms. The difficulty of reaching and enforcing a joint profit maximizing result by a group of oligopolists, which involves interfirm contracting of an implicit kind, is contrasted with the ease of achieving such an outcome by a dominant firm (monopolist).

Dominant firms enjoy collusive advantages over oligopolists in three significant respects. For one thing, the dominant firm does not have to write a complex interfirm "contract" in which future contingencies are identified and appropriate adaptations thereto are devised. Rather, the dominant firm can employ an adaptive, sequential (intrafirm) decision-making procedure, which economizes greatly on bounded rationality demands, without exposing itself to the risks of contractual incompleteness which confront autonomous parties to a horizontal interfirm agreement. Second, adaptation within a firm (in contrast to that between firms) is promoted by the more complete development of efficient, albeit often informal, communication codes and an associated trust relationship between the parties. Third, as between semiautonomous operating divisions in a dominant firm and fully autonomous oligopolists, the incentives to cheat are less and the machinery for penalizing defectors is more efficacious in the dominant firm. *MH* accordingly concludes that the net benefits of transforming oligopolies to monoplies, thereby to activate voice, are doubtful. Indeed, the opposite policy recommendation is reached: subject to the condition that a sacrifice of scale economies is not implied, dissolution of dominant firms into rivalrous oligopoly warrants serious attention.

### 2. *Capital Markets*

*EVL* pays scant attention to competition in the capital market. To the extent that it does, such competition is characterized as effete. *MH*, by contrast, regards competition in the capital market as integral and points up the symbiotic relations that have developed between internal organization (which, if of the appropriate kind, can usefully be regarded as a miniature capital market) and the capital market itself. Conglomerate organization is interpreted in these terms.

The argument here is that, where competition in the product market is weak or lacking, capital market competition for managerial control over corporate assets has an important role to play. The hierarchial transformation of firms from traditional, functionally organized entities to divisionalized structures permits an *internal* resource allocation competition to be set up which facilitates the assignment of capital to high yield purposes. Although the divisionalized firm's investment opportunities are limited, its *ex ante* knowledge with respect to each is very deep and its *ex post* ability to monitor investments is great. Both adverse selection and moral hazard problems associated with conventional capital market operations are thereby attenuated.

### 3. *Command Economies*

Command economy issues of two types should be distinguished. One entails the polity; the other concerns the organization of firms. *EVL* explicitly addresses the former but not the latter. The *MH* treatment of firms in an enterprise system has a bearing on the organization of firms in a command economy, but the issues are not expressly developed.

*EVL* notes with respect to the polity in command economies that "With exit-competition playing a much smaller role in the Soviet economy than in the market economies of the West, it was found necessary to give voice a more prominent role" (p. 34). It offers no corresponding advice, however, for the organization of firms.

Although *MH* is concerned throughout with the organization of an enterprise system, the organization form issues which are addressed in this context apply more generally. As compared with conventional industrial organization, where internal organization is neglected and concern over market organization predominates, *MH* argues that internal organization profoundly matters. To shift a transaction out of the market and into the firm without simultaneously examining the way in which such a transaction will be subject to internal incentives and controls is at best hazardous and can be foolhardy.

The limits of hierarchies in both vertical integration and firm size respects are relevant in this connection. Both capitalist and socialist firms experience expenditure distortions as the incremental transaction is shifted out of the market and into the firm. Unsurprisingly, the extent to which these develop is a function of organization form.

### B. *Public Policy*

Quality issues are of major concern throughout *EVL*. These appear in connection with the Nigerian railway problem and in the discussions of oligopoly and connoisseur goods.[6] And the connoisseur goods discussion leads into general system concerns that the possible failure of voice to be mobilized in requisite degree will result in deterioration with respect to "a number of essential services largely defining what has come to be called the 'quality of life' " (p. 53).

The strategy of exit-blocking so as to induce voice, thereby to impose connoisseur preferences on the remainder of the population, can be addressed in two parts. First, in what circumstances, if any, can it be justified in simple efficiency terms? Second, where these conditions do not obtain, what *prima facie* indicators ought to be satisfied before exit-blocking is seriously entertained? This latter is the more difficult issue and is not addressed here. I nevertheless think it instructive that exit-blocking need not turn, always or entirely, on complex value judgments of a nonefficiency kind.

Thus assume that the only significant differences between two consumer groups is that one has greater time, inclination, or capacity to evaluate alternatives in a discriminating way. Imposing the choice of the more discriminating group on the other can then be regarded, I submit, as an efficient outcome in conventional specialization of labor terms. (That it does not obtain "naturally"—i.e., without public policy assistance—is presumably explained by frictions in the system.) In circumstances, however, where the groups arguably differ in preference and other respects, exit-blocking is a political decision with deeper ramifications. The possibility of supplying immobile or disadvantaged sectors of the population with wherewithal, so they could exercise exit, rather than constrain already mobile segments ought also be considered as a remedial measure.

*MH* raises quality considerations mainly in conjunction with the discussion of atmosphere and then in only a limited way. To the extent that alternative contracting modes differ in atmospheric respects and these affect the quality of life on the job, neglect of these will result in suboptimization. Put differently attitudinal nonsepara-

[6] Quality considerations also come up in discussions of how the quality of government is affected by loyalty.

bilities are neglected only at the risk of transactional inefficiency.

The public policy thrust of *MH* is mainly with respect to antitrust enforcement. A revision of antitrust thinking is urged in which transaction cost considerations are expressly taken into account. More discriminating policies with respect to vertical integration, conglomerate organization, dominant firm industries, and oligopolies are proposed.

Also, *MH* conjectures that problems of devising market alternatives to regulation for natural monopoly industries are apt to be more severe than the advocates of franchise bidding (Harold Demsetz 1968; R. A. Posner 1972) have indicated. Subsequent application of the contracting approach to the franchise bidding issue supports the conjecture (Williamson 1976).

## IV. Concluding Remarks

As indicated at the outset, the paper's emphasis on the economics of internal organization biases the discussion in favor of the *MH* approach. Despite this bias, it is apparent that *EVL* has many interesting insights to contribute to the study of economic organization. In many respects, *EVL* and *MH* are complements to each other as well as complements to received micro theory.

Of special significance in this respect is the fact that *EVL* is mainly concerned with community objectives of a quality of life kind while *MH* is more narrowly concerned with transactional efficiency. Inasmuch as both are vital in the design of any high performance system, addressing complex economic problems in both the more aggregate language of exit and voice as well as in the more microanalytic contracting language of markets and hierarchies is apt often to be fruitful.

A comparative institutional approach to the study of economic phenomena is employed by both. The alternative to markets that is emphasized in *EVL* is voice. Shifting transactions out of the market into the firm (in labor, intermediate product, and capital market respects) is studied in *MH*.

The object of *EVL* is to mitigate the dysfunctional consequences that are associated with the unrestrained exercise of micromotives. Blocking or inhibiting exit, thereby to *mobilize voice* and achieve a *collectively superior* outcome is the recurring theme.

The object of *MH* is to evaluate organizing modes in terms of transactional efficiency, where this is judged comparatively. Institutions are assessed with respect to their properties for *economizing on bounded rationality and attenuating opportunism. Ceteris paribus*, those market and internal structures that realize transactional economies in these respects will display superior survival qualities.

## REFERENCES

**K. J. Arrow**, "The Organization of Economic Activity," in *The Analysis and Evaluation of Public Expenditure: The PPB System*, Joint Economic Committee, 91st Congress, 1st ses., 1969, 59–73.

———, *The Limits of Organization*, New York 1974.

**R. Coase**, "The Nature of the Firm," *Economica N.S.*, 1937, *4*, 386–405; reprinted in G. J. Stigler and K. E. Boulding (eds.), *Readings in Price Theory*, Homewood 1952.

**J. R. Commons**, *Institutional Economics*, New York 1934.

**R. M. Cyert and J. G. March**, *A Behavioral Theory of the Firm*, Englewood Cliffs 1963.

**H. Demsetz**, "Why Regulate Utilities?" *J. Law Econ.*, Apr. 1968, *11*, 55–66.

**P. Doeringer and M. Piore**, *Internal Labor Markets and Manpower Analysis*, Boston 1971.

**I. Goffman**, *Strategic Interaction*, Philadelphia 1969.

**F. A. Hayek**, "The Use of Knowledge in Society," *Amer. Econ. Rev.*, Sept. 1945, *35*, 519–30.

A. O. Hirschman, *Exit, Voice and Loyalty*, Cambridge 1970.

H. Leibenstein, "Allocative Efficiency vs. 'X-Efficiency'," *Amer. Econ. Rev.*, June 1966, *56*, 392–415.

R. R. Nelson and M. Krashinsky, "Public Control and Economic Organization of Day-Care for Young Children," *Public Policy*, Winter 1974, *22*, 53–75.

R. A. Posner, "The Appropriate Scope of Regulation in the Cable Television Industry," *Bell J. Econ.*, Spring 1972, *3*, 98–129.

T. C. Schelling, "On the Ecology of Micromotives," *Public Interest*, Fall 1971, *25*, 61–98.

H. A. Simon, *Administrative Behavior*, New York 1957.

R. M. Titmuss, *The Gift Relationship: From Blood to Social Policy*, New York 1971.

O. E. Williamson, "The Economics of Antitrust: Transaction Cost Considerations," *Univ. of Pennsylvania Law Rev.*, June 1974(a), *122*, 1439–96.

———, "Exit and Voice: Some Implications for the Study of the Modern Corporation," *Social Science Information*, 1974(b).

———, *Markets and Hierarchies: Analysis and Antitrust Implications*, New York 1975.

———, "Franchise Bidding for Natural Monopolies—in General and with Respect to CATV," *Bell J. Econ.*, Spring 1976, *7*.

———, M. Wachter and J. Harris, "Understanding the Employment Relation: The Analysis of Idiosyncratic Exchange," *Bell J. Econ.*, Spring 1975, *6*, 250–80.

# Consolidation or Diversity: Choices in the Structure of Urban Governance

By Dennis R. Young*

In very large measure, issues of structure in metropolitan area governance can be characterized by a basic organization choice: consolidation or diversity. Along several dimensions, a basic policy variable is whether public decisions are to be made by a few large organizations or units of government, or many small ones. Two key dimensions are *spatial* and *vertical* organization.

*Spatial organization* refers to the drawing of political boundary lines over urban space, dividing population groups into separate political constituencies. The issue for urban governance is how such boundaries should be drawn—specifically, whether they should encompass large or small areas, and small, homogenous or large, diverse population groups. *Vertical organization* refers to the possible separation of the "collective demand articulation" from the "supply" aspects of public service provision. Governments often have the options of contracting to, or subsidizing, private or other public suppliers rather than producing services "in house." To what extent should the supply and demand aspects of government be separated, and supply arrangements diversified?

A number of contributors to the field of political economy have provided insights into the resolution of these organizational questions. Albert O. Hirschman's (1970) theory of "exit and voice" is especially germane, as "consolidation" usually implies an organizational structure more dependent on the nonmarket forces of "voice," while diversification often implies increased reliance on market-like forces of "exit."

* Associate Professor, W. Averell Harriman College for Urban & Policy Sciences, State University of New York at Stony Brook.

## I. Spatial Organization

Two key principles underly much of the existing economic theory applying to delineation of political boundary lines over urban space. The first principle, advanced in a classic paper by Charles Tiebout (1956) is that urban areas can be organized into a heterogeneous, polycentric system of many small communities whose individual populations have relatively homogeneous tastes for public goods. Such a mode of organization is said to allow citizens to locate in communities whose public service packages and tax burdens conform most closely to their individual preferences. In Tiebout's model, an equilibrium is developed via citizen mobility, under which local public goods are allocated in an approximately optimal fashion, analogous to the functioning of the marketplace for private goods.

Tiebout's theory assumes the absence of interjurisdictional externalities. Such external effects between communities not only upset the optimal allocation of resources of the Tiebout model, but of course represent a key problem of urban governance. Political economists have offered two approaches to the resolution

of intercommunity externality problems. Those in the Tiebout tradition of favoring polycentric governance have stressed the use of intergovernmental agreements to ameliorate interjurisdictional externalities. This simply means that adjacent communities whose activities impose benefits or costs on one another can bargain in order to reach a level of externality production that would be socially optimal. This solution seems desirable except for the possibility that external effects may be widespread and may involve many different jurisdictions. Under such circumstances, the possibility of a bargain becomes remote, as the decision-making costs of reaching agreement among affected communities becomes very large.

The second mode offered by political economists for the solution of interjurisdictional externality problems is internalization via the adjustment of political boundary lines so that external effects are fully encompassed by single political jurisdictions. This is Mancur Olson's (1969) principle of "fiscal equivalence," under which a given political constituency would face both the full costs and the full benefits of a particular public service or action. Olson notes the delicacy of drawing boundaries via this principle. In particular, boundary lines encompassing too small an area will fail to fully internalize external effects. On the other hand, boundaries drawn too widely risk the danger of "internality." In this case, relatively few people in a confined part of a jurisdiction are affected by a particular externality or (proposed) public good, but since the whole jurisdiction is required to support a collective action, a satisfactory resolution may not take place. Thus, there is the danger of substituting "political externalities" for interjurisdictional externalities, by moving from boundaries that are too narrow to boundaries that are too wide.

The alternative of a federal structure of government presents, at least in theory, an attractive compromise to the Tiebout polycentric system on the one hand, and complete consolidation on the other hand. Metropolitan federalism offers the possibility of retaining the political identity of communities at the local level while allowing a higher level of government to ameliorate intercommunity external effects. But the compromise is certainly not without its own problems. Clearly an all-encompassing layer of government at the metropolitan level will face problems similar to "internality." That is, controlling coalitions of local communities may impose political externalities on other localities in the metropolitan area.

Hirschman's theory of "exit and voice" adds a few new wrinkles to the foregoing traditional political economy arguments. For example, Hirschman (1974) observes that reliance on the exit option is far too heavily emphasized by the Tiebout theory. In the first place, even where mobility is high, people move for a mixture of motives, not necessarily dominated by public goods considerations. Second, the level of mobility, even among high income groups is restrained. Hirschman describes the phenomenon of "exit fatigue" that quickly sets in with respect to moving one's residence. However, limitations on exit do not necessarily contradict the efficacy of a system of many small, homogeneous, self-governing communities. In particular, the mechanism of voice itself may be more effective when applied in the context of smaller, homogeneous rather than large heterogeneous communities. This, in fact, seems more important than the exit option, in explaining the empirical findings of researchers such as Elinor Ostrom and Gordon Whitaker (1973), Robert Bish (1971), and Brian Stipak (1974), that small, homogeneous communities seem to be more successful in achieving desired

levels and quality of services if they are politically self-contained than if they are part of a larger jurisdiction.

There are several reasons for the possibly increased effectiveness of voice in the context of smaller, more homogeneous communities. First, is the phenomenon of internality. Political externalities can be imposed on non-self-governing communities contained within large political jurisdictions, but not on self-governing communities. In addition, there is the likelihood that citizens will participate (use voice) more in smaller jurisdictions where they have a higher probability of impact and can see what effects they are having. Such an hypothesis is consistent with Olson's (1971) theory of collective action, which Roland McKean (1974) uses to explain the failure of citizens to participate in the political process. It also seems consistent with Hirschman's (1974) observation that exercise of voice can be sometimes viewed as a pleasurable or fulfilling experience. In particular, we would surmise that such fulfillment would be enhanced by the greater expected impact of voice in smaller jurisdictions.

Other aspects of the exit-voice framework, however, seem to contradict the efficacy of a system of many small jurisdictions, especially where mobility *differs* substantially among population groups. In particular, a polycentric system implies the presence of choice in selecting one's jurisdiction of residence, and the *potential* use of exit as an ultimate, if not immediate, response to deterioration of the services in one's present community. But, as Hirschman notes, exit will first be used by mobile, high income consumers who are more sensitive to quality than cost. These citizens will move to more expensive (suburban) jurisdictions leaving behind (and often effectively locking in) people who are less mobile, less quality-conscious, and less willing or able to utilize voice to repair the quality deterioration that has taken place. For these and other reasons, communities become segregated into those with high quality, high cost and low quality, low cost programs.

Exit-voice theory suggests that it may be preferable to eliminate the exit option in such a system, i.e., to lock quality-conscious consumers into the same jurisdictions as their less quality-conscious compatriots. This could be effected by replacing the polycentric arrangement with a consolidated metropolitan government. Under such a consolidation, quality-conscious consumers, having nowhere to exit, would presumably use voice more effectively in checking general deteriorations in service.

But consolidation is no panacea. The key considerations appear to be income and political control. As long as some people are poor, they will be more price than quality conscious. As long as middle income groups are numerous and powerful, they will maintain political control and skew the allocation of public resources toward their own ends, irrespective of the governing arrangement. Middle and upper income groups operating within a polycentric governmental structure, create suburban tax havens through zoning, to eliminate the transfer of resources to the support of low income communities. Alternatively, under a consolidated structure, middle and upper income groups utilize the political process to effect the same ends within a jurisdiction. It is simply testimony to the relative inefficiency of the latter or the subtlety of the former in inhibiting resource transfers, that suburban jurisdictions have opposed consolidation.

We can speculate that effective resolution of urban governmental spatial structure may require multiple independent communities controlled by different population groups. Alterations of the price-

quality trade-offs would have to be accomplished by structural reform in the form of income redistributive arrangements (e.g., revenue sharing) at the metropolitan level. In this manner, the effect of local control would be to strengthen voice of each (relatively homogeneous) citizenry, while the effect of income redistribution would be to redress the price-quality trade-off of lower income communities.

The issue of intercommunity externalities is placed within the context of exit-voice theory by noting that by definition, there is but limited exit possible from widespread external effects. In such a circumstance, a voice-oriented institutional framework seems required. In the urban environment, this suggests areawide units of government overarching local communities, for services such as transportation, air and water quality, waste disposal, and energy distribution. But in view of the aforementioned difficulties in using voice in large heterogeneous governmental jurisdictions, how can this level of government be successfully structured? The fundamental problems persist: minorities will be neglected by virtue of their submergence in a majority sea, while participation will be inhibited by the sheer size of the jurisdiction. These effects may be minimized but not eliminated by keeping the size of areawide jurisdictions to smallest units consistent with substantially internalizing relevant externalities.

Given the existence of public functions involving extensive areawide effects, there seems no easy way to circumvent these problems in the implementation of voice. Areawide levels of government are necessary, and necessarily imperfect.

## II. Vertical Organization

There are various options for governmental units to ensure that public services are supplied to their jurisdictions: governments can produce services themselves, contract with other producers, or allow citizens to use the marketplace under arrangements for possible subsidy and regulation. Robert Warren (1964), Bish (1971) and others point out that the foregoing "supply response" aspect of public goods provision is conceptually distinct from the "demand articulation" function of government, and should be organized by different criteria. Thus, it is argued that governmental units need not be large enough to fully exploit scale economies, or to produce every variety of differentiated product, for they can always contract with efficiently sized or specialized (private or other public) suppliers.

Furthermore, it is argued that governmental units should focus solely on demand articulation. Political economists of the polycentric school reason that a system featuring local communities organized as "consumer cooperatives," purchasing from multiple suppliers competing for contracts with these cooperatives, is efficient both in a production sense—controlling the cost and quality of services, and an allocation sense—matching service outputs to local preferences. But this conclusion, which is heavily dependent on the effectiveness of exit (by communities) in the markets for public services, would seem to vary with the nature of particular public services. For example, Oliver Williamson (1973) notes that "contractual incompleteness"—the inability or costliness of exactly specifying the desired nature of a service or the expected behavior of a contractor under all possible contingencies—can mar the desirability of contracting relative to "in-house" supply. In simplest terms, contracts appear to work best in areas where all important aspects of the output are quantifiable, where the technology is known and costs are predictable, and where a degree of competition can be assured.

A basic option alternative to formal

contracting or in-house supply, is to sanction private supply under some type of franchise or regulatory scheme which may involve rate setting, public subsidy, or the enforcement of various operating rules and guidelines. There is, of course, a wide spectrum of practices in this category ranging from regulation of monopolistic utilities, through administration of voucher programs to finance consumer use of competitively provided services. Arguments for privately supplied, publicly regulated services stress the motivational benefits to suppliers of self-financing and (in some cases) competition, as well as the ability of individual consumers to select the specific quantities desired at a given price.

The exit-voice framework provides an interesting perspective on these questions of consolidated or diversified, public or private supply. In particular:

Government in-house supply depends for its discipline directly on citizen voice expressed through the local political process. This voice is inevitably inexact and "noisy" as it must be integrated through channels of complaint and voting.

Government contracting combines exit and voice. Voice is directed by citizens, through the political process, to governmental overseers. Those overseers exercise both exit and voice in attempting to discipline contractors. A presumed advantage of contracting, as already noted, is that it strengthens citizens' voice by converting local government into a kind of consumer cooperative whose interests are aligned with the citizens, rather than a supplier whose interests may be at odds with those of the citizenry.

Government regulation of private monopoly provides for very weak forms of both exit and voice. Exit is hampered by the paucity of alternatives and the essentialness of services. Voice is muffled by the usual insulation of regulators from the political process, and the unequal strengths with which supplier and consumer interests are heard by regulators. On the other hand, regulation carries within its framework of definition the potential for the exit and voice strengths of the contracting arrangement. In particular, voice can be strengthened by bringing regulatory commissions into the general governance structure. Exit can be strengthened by eliminating the permanence of franchises and introducing precise performance criteria upon which rate or subsidy increases, and franchise renewals, would be granted.

Use of competitive private supply takes greatest advantage of control via exit. Here the consumer wields direct power via his choice to patronize one producer or another. Interestingly, competitive supply can also exhibit the most direct manifestation of consumer voice. For under this arrangement, consumers are more likely to articulate their concerns directly to suppliers before exercising exit. Furthermore, it is under this arrangement that suppliers are most likely to listen to these same consumers (precisely because of the threat of exit).

Beyond these general observations, particular characteristics of services influence the efficacy of exit and voice, and the desirability of one or the other of the foregoing supply arrangements. For example:

—economies of scale may preclude or discourage the provision of choice among several suppliers; hence, exit via competitive supply arrangements may be unrealistic for such services as telephone, electricity, rail transport, water supply, and sewers.

—the difficulty of precisely specifying the desired nature of service output or supplier behavior in all contingencies can make exit a sterile means of control. "Information rich" voice (dialogue between consumers and producers) is needed to reach a satisfactory resolution, and avoid

the risks of strategic behavior under misguided contract specifications. Thus, in-house supply may be preferred to contracting, and monopoly, or costly exit, may be preferred to competitive arrangements for this reason. An example here is elementary schooling whose output is sufficiently complex and ephemeral so that contracts cannot be completely specified and parental exit does not neatly convey the reasons for dissatisfaction.

—some services pose natural exit barriers, i.e., high personal costs for changing one's source of supply, which reduce the efficacy of competitive arrangements by taking the edge off the threat of exit. Examples include day care centers and nursing homes where changes are disruptive of the lives of clientele; or housing where exit involves the high dollar costs of moving.

—for services which Hirschman describes as "connoisseur goods," exit may be limited to those who can afford more expensive private alternatives. Such is the case with education, health care, and housing. In this circumstance, the danger is that "quality-conscious" consumers will exit, leaving behind those consumers who cannot exercise exit nor effectively use voice. Hirschman's analysis suggests that barriers to exit (or abolition of alternatives) would help prevent quality deterioration by keeping the most effective wielders of voice locked in. On the surface, this hypothesis appears to argue for monopolistic arrangements which (hopefully) channel voice effectively. Alternatively, it may well be that correcting income balances through subsidy can do much to void the problems of connoisseur goods. A combination of corrective subsidy plus consumer information might ease the segregation of quality and nonquality-conscious consumers and reduce the correlation of this bifurcation with income and minority groups (just as local revenue sharing might do the same for communities as a whole).

—some services have no viable consumer groups that can competently exercise exit (and perhaps voice also). Services provided to children, the ill or infirm, or criminal offenders, clearly preclude exit as a viable option at the consumer level. Hence, the options are seemingly reduced to in-house government supply or contracting. But these arrangements are also hindered in these circumstances by the absence of a strong advocacy group to whom government would be responsive. A possible circumvention of this problem, suggested by this author (1974), is the use of "proxy consumers" to make exit choices and exercise voice on behalf of individual consumers. For children in school, or elderly persons in nursing homes, members of the immediate family often serve this purpose. But even in such areas as criminal corrections or foster care, judges, probation officers, social workers and lawyers potentially fit this role.

Given the foregoing variety of service characteristics which affect the efficacy of exit and voice, we should expect that different services can be best organized vertically in different ways. For example, if exit is hampered by exit costs, vague service character, or absence of a viable consumer, then contracts might be avoided in favor of in-house control, and consumer's choice must be supplemented by voice-responsive governmental supervision of some kind. Or if exit is effective relative to voice, i.e., where service character is well defined, consumers are viable, and exit costs low, then competitive or contract arrangements could be emphasized.

In contrast to the discussion of spatial organization of urban government, the economist's bias towards exit as the principle means of regulation, appears to serve him better in the domain of vertical organization of public services. While there

is a considerable variety of practices in the United States at the present time, it seems safe to propose that voice-dependent modes of supply organization are emphasized too heavily for local public services. Recent experiences in education, sanitation, even fire protection, indicate possibly greater useful exploitation of exit, through contracting and competition, to replace stagnant forms of bureaucratic and regulated monopoly supply. (See E. S. Savas, 1974.)

### III. Conclusion

Consideration of the various problems in each organization dimension of urban governance reveals no uniform preference for voice-based (consolidated) or exit-based (diversified) modes. In the extremes, complete spatial and vertical integration would bring out in full bloom the circumstances under which voice is dysfunctional, while full diversification along these dimensions would do the same for exit. The choices in between are numerous, varied, interdependent, and dependent upon the character of individual services, and clientele populations. A review of this discussion reveals various circumstances where exit seems to work best and where exit fails, and where voice seems to work best and where voice fails.

Economists are fond of the term "market failure" which is taken to mean the failure of market arrangements to bring about an optimal allocation of goods and services. Using Hirschman's terminology we can take the liberty of characterizing market failure more broadly, as the circumstances in which the use of exit, or exit-based organizational arrangements, fail to maintain organizational performance. Such a definition would extend the classical one to include not only the conventional conundrums of externalities and economies of scale, but also the various other factors discussed here such as exit costs and complexity of output. Given the characterization of market failure as failure of the exit mechanism, it is but a short step to defining a complementary concept—"nonmarket failure." Nonmarket failure would be the circumstances in which the use of voice, or voice-based organizational arrangements, fail to maintain organizational performance.

The twin concepts of market failure and nonmarket failure bring to mind two immediate ideas seemingly worthy of further development. First, just as we can envision ideal market conditions under exit based arrangements (e.g., perfect competition, with all its concommitant assumptions), we can also attempt to visualize ideal circumstances for voice-based arrangements—perhaps where self-governing communities or organizations are composed of small homogeneous units. Given these "ideals," one approach to organizational reform might be to attempt to approximate these ideals in practice.

A second outgrowth of the duel concepts of market failure and nonmarket failure is not so sanguine, however. A review of the various circumstances under which exit fails and voice fails reveals that voice does not necessarily succeed where exit fails. For example, overcoming the exit failure due to areawide externalities in urban areas requires a large, heterogeneous voice-based metropolitan government structure which is also subject to voice failure. Thus, for a given circumstance, market failure and nonmarket failure are *simultaneously* possible!

It is likely therefore that there is a core of public services and other economic activities for which no perfectly satisfactory organizational solutions may be available. While this appears to be a very pessimistic conclusion, it may have its redeeming virtues. For it may serve to warn reformers of various persuasions, both econ-

omists and political scientists, that "the grass is not necessarily greener on the other side." Some areas of economic activity at least, appear to be necessarily organized in an imperfect way.

## REFERENCES


R. Bish, *The Public Economy of Metropolitan Areas,* Chicago 1971.

J. Buchanan and G. Tullock, *The Calculus of Consent,* Michigan 1962.

A. O. Hirschman, *Exit, Voice and Loyalty,* Cambridge 1970.

———, "Exit, Voice, and Loyalty: Further Reflections and a Survey of Recent Contributions," *Social Science Information,* Feb. 1974, *13,* 7–26.

M. Olson, "The Principle of 'Fiscal Equivalence': The Division of Responsibilities Among Different Levels of Government," *Amer. Econ. Rev. Proc.,* May 1969, *59,* 479–87.

———, *The Logic of Collective Action,* Cambridge 1971.

E. Ostrom and G. Whittaker, "Does Local Community Control of Police Make a Difference: Some Preliminary Findings," *Amer. J. Polit. Sci.,* Feb. 1973, *17,* 48–76.

R. McKean, "Collective Choice," Ch. 5 in James W. McKie (ed), *Social Responsibility and the Business Predicament,* Washington 1974.

E. S. Savas, "Municipal Monopolies Versus Competition in Delivering Urban Services," Ch. 15 in Willis D. Hawley and David Rogers (ed.), *Improving the Quality of Urban Management,* Beverly Hills 1974.

B. Stipak, "Citizens Evaluations of Municipal Services in Los Angeles County," pap. no. 194, Institute of Government and Public Affairs, Univ. of Calif. at Los Angeles, Sept. 1974.

C. Tiebout, "A Pure Theory of Local Expenditures," *J. Polit. Econ.,* Oct. 1956, *64,* 416–24.

R. Warren, "A Municipal Services Market Model of Metropolitan Organization," *J. Amer. Inst. Planners,* Aug. 1964, *30,* 193–204.

O. E. Williamson, "Markets and Hierarchies: Some Elementary Considerations," *Amer. Econ. Rev. Proc.,* May 1973, *63,* 316–25.

D. R. Young, "Exit and Voice in the Organization of Public Services," *Social Science Information,* June 1974, *13,* 49–65.

# DISCUSSION

ALBERT O. HIRSCHMAN, Institute for Advanced Study: In the venerable folklore of traditional economics political action in the economic sphere has typically been considered as noxious interference with the totally beneficent market mechanism The equally venerable opposing point of view castigates the Invisible Hand as the Blind Forces of the Market—note the clever replacement of Adam Smith's metaphor by one that uses a closely related image—and instead advocates planning by a political authority that, in turn, is assumed to be totally beneficent. A principal intent of my work on exit and voice, considerably bolstered by all three papers presented at this session, was to show that there is a wide range of economic processes for the efficient unfolding of which both individual, economic action (via exit) and participatory, political action (via voice) have important constructive roles to play. The novelty of the theory for the economist—apt to infuriate or to disconcert both market and planning enthusiasts—consists in its stress on voice not as a substitute for the market nor as a restraint on it in a few well defined situations, but as another generally available mechanism that, like the market, has its strengths and weaknesses, its successes and failures. There is also a stress on the possibility of unstable equilibria between exit and voice as one drives out the other and on the lack of once-and-for-all solutions through optimal mixes of exit and voice. In other words, the new approach does not satisfy our craving for equilibrium, harmony and final repose. If it has met, nevertheless, with the beginning of a response from economists, this is probably because of the new or renewed emphasis on such phenomena as ignorance and the cost of information, transaction costs, bounded rationality, X-efficiency, and even altruism. The possibility of "market failure" is today no longer limited—as was taught not long ago by received doctrine—to the presence of externalities, and I welcome Dennis Young's proposal to redefine this concept broadly in terms of situations where exit doesn't do a good job of stirring up management and of restoring efficiency. In this perspective the availability of another mechanism—voice—should not be viewed as a threatening rival, but as a welcome resource whose potential contribution to a containment of our multiple troubles deserves to be closely studied. For example, while looking at the provision of medical services precisely in this manner, Carl Stevens has recently proposed to complement consumer sovereignty—that is, exit—by *consumer participation* and to limit professional producer sovereignty by *lay management constraints*. These terms are useful modulations of the perhaps excessively compact "voice."

My comments, including some further thoughts will be in four parts.

1. In all three papers, when voice scores over exit this is primarily because it is information-rich whereas exit conveys little beyond the fact of restlessness or restiveness on the part of the consumer, customer, employee or whoever else one has an economic relation with. Take Freeman's paper on labor and the employment relationship: here dissatisfaction arises from a mixture of motives that includes unhappiness about a whole array of working conditions no less than the desire for higher wages and benefits. Unions are more efficient in communicating information about this complex mixture of complaints than the personnel turnover rate. To make this point even stronger I would add that the *content of discontent* is apt to change over time: exit always takes the same form no matter what the complaint; but voice can articulate newly arising demands—often, to be sure, with a considerable lag as the union leadership takes its own time to understand and adopt the new "outlandish" grievances of its members.

According to Williamson, markets are not working properly whenever the parties are involved in a relationship which it is difficult to specify fully in writing because of uncertainty and complexity or which lends itself peculiarly to the display of bargaining behavior. In these cases the substitution of hierarchy for market relationships between the parties is indicated. In my language hierarchy can be considered as a special variety of insti-

tutionalized voice and here is therefore another case where voice wins out over exit because of the fullness and frankness of information which it carries.

Finally, Dennis Young uses the same criterion in favoring a large admixture of voice for the organization of certain public services. He mentions primarily situations in which exit is costly and disruptive to the consumers of the services, as in the case of day-care centers or nursing homes. In my view, these are also frequently situations in which the producers themselves are still groping for satisfactory ways of doing a job which they have taken on because there was a demand for it. In such increasingly important cases of *producer ignorance*, information-rich voice on the part of consumers or of Young's "proxy consumers" is far more helpful than the blankness of exit.

2. That voice conveys more information than exit is one facet of the polarity which is highly important for economic applications, and which I had neglected in my book. Another is the fact that voice, to the extent that it is political action or action in the public interest, is liable to escape from the fetters of the benefit-cost calculus and can therefore suddenly gain an unexpected edge over silent, self-regarding exit.

Young notes quite correctly one dilemma in the design of local government: the larger the territorial unit encompassed by municipal government, or by a subdivision of that government such as a school district, the less feasible will it be for the individual citizen to practice exit and the more important is it therefore for voice to function actively and intelligently; yet voice is presumably more apt to become mobilized when the area to which it extends is none too large. Decentralization of public schools and of public services has been advocated precisely for the purpose of facilitating and activating citizen participation.

It may be questioned, however, whether the propensity to voice is likely to increase *indefinitely* as the territorial unit within which voice is raised shrinks in size. Voice can become an end in itself, and its exercise can become confused with the attainment of its objective. As soon as that happens its *cost* (in time spent, for example) can measure instead the *satisfaction* or *benefit* received from its exercise or "consumption." For this mutation of cost into benefit to occur it is necessary that the exercise of voice be felt as something beyond the many activities that are primarily self-regarding. Though originally prompted by personal concerns, voice becomes an enjoyable, exhilarating experience when it is also action in the public interest, sometimes just because it is felt as a release from the unremitting pursuit of purely self-regarding activity. But voice stands to lose this public interest dimension if the area within and on behalf of which it is raised shrinks unduly: it then becomes essentially self-regarding like exit and at that point its cost will once again be rigidly computed. Here is perhaps a simple explanation of the consistently higher voter turnout in national as compared to local elections. In local elections, the public interest dimension is less obvious and the cost of voting tends to be computed and related to its conceivable benefits; national elections, on the other hand, partake of the character of public celebrations and even of the traditional "feast of fools": the lowly citizen is transformed into the sovereign—though only for one day as Rousseau lamented—and he enjoys himself thoroughly in the process.

3. What has just been said explains much about the economist's preference for exit: voice is a far more complex, less predictable mechanism. In my book I stressed the similarities between the ways in which exit and voice operate. I overlooked or failed to stress sufficiently certain characteristic differences. He who voices remains within the organization, maintains a relationship with it. Hence the organization and its management can do things to him and in particular can treat him, because he voices, differently from the way in which it treats the other members.

It can treat him *worse*, punish him for daring to voice. There may be *retaliation* and *reprisals* against someone who criticizes an organization, but remains within its reach. For that reason, it has been argued by Birch that the availability of a fast getaway (exit) is a precondition for a good volume of voice. I agree but would also point out that a number

of social mechanisms have been evolved to make voice retaliation-proof even in the absence of exit. Examples are the secret vote, the Ombudsman institution which make it possible for individuals in a bureaucratic organization to complain outside of hierarchial channels, and trade union bargaining as described by Freeman.

Another hazard that exists for voice far more than for exit is the exact opposite of retaliation: instead of being penalized, he who voices is singled out for *special favors*. By giving the complainant preferential treatment the delinquent organization seeks to still his voice and to buy him off: in this manner, it may once again avoid having to improve the *general* quality of its performance. This managerial strategy is present when firms extend "goldplated service" to their more important customers; it is also often practiced by administrative agencies rationing out goods and services to applicants according to nonmarket criteria.

This strategy of countering voice is employed not only for the purpose of appeasing important individual customers; it can be particularly useful to management in dealing with voice that takes collective form, as in the case of trade unions, consumer organizations and the like. Special treatment and favors extended to the leaders of such organizations could have results that range all the way from a slight reduction in militancy to outright corruption. This is one of the major risks to which the effectiveness of collective organized voice is exposed.

4. It was a general underlying assumption of my book that both forms of activism, voice and exit, would, if successful, achieve a benefit (a positive externality) for the nonactivists. This assumption would hold whenever a firm turns out a product (or the government a policy) the tastes for which are so constrained that any improvement in quality or performance undertaken by management in response to exit and voice is sensed as a positive event by everyone, activists as well as nonactivists. What happens when this constraint is relaxed? Suppose that different consumer-members have different ideas about what sort of improvements are needed and further that the ideas and tastes of the activists differ systematically from those of the nonactivists. To the extent that it is successful, the voice of the activists will then cause the quality of the product or policy to vary in such a fashion that benefits are bestowed primarily or exclusively on them. The reason is simple: voice is information-rich and is able to give precise instructions to management. Note that those who voice will here receive special benefits without any conscious buying-off activity on the part of the organization.

This constellation has implications that differ considerably from my previous model. I had argued that, when exit is available and loyalty weak, deterioration would make for rapid desertion of the potentially most influential carriers of voice and would thereby cause further deterioration of organizations that are so constituted as to respond more to voice than to exit. I still hold that this model corresponds to one set of important situations in the real world. But to the extent that the just noted conditions prevail, the outlines of another set with almost opposite characteristics appear. It can happen that an institution originally or nominally set up to service a wide group comes to cater predominantly to the wishes and tastes of an articulate minority or oligarchy within that group. If exit is available at all under these circumstances, it could then become the weapon that will typically be wielded by the "silent majority." In this perspective, the articulate few would develop a preference for voice whereas the comparative advantage of the underarticulate mass would lie in exit; exit might in fact be the only means of defense of the voiceless not so much against deterioration as against voice-induced quality changes that are not in their own interest. (I am indebted to Pierre Bourdieu and David Riesman for raising this point.)

Actually examples of this reversal of roles of exit and voice are not easy to find. While it is commonplace for voice to function as an instrument of the privileged and for the privileged, exit has not often been used to provide an avenue of self-defense for the voiceless. An important case in which exit did approximate this function was during the

nineteenth and early twentieth centuries in connection with mass emigration from Europe to the New World. But, on the whole, exit has not been readily available to the voiceless masses—they have had to assert themselves the hard way through passive resistance, revolts, and other attempts at voice. It has not often been the case that the advantage of the privileged with respect to voice is offset by a similar advantage the underprivileged might have with respect to exit. To the contrary, because of a number of technological and institutional innovations, the privileged have in recent times compounded their traditional superiority in voice with a remarkable prowess in exit—witness the flight to the suburbs, capital flight, the multinational corporation and the brain drain. In emphasizing the disruptive consequences of such exits and in looking to more voice for a remedy, I may well not have given enough attention to the possibility that by using voice instead of exit the articulate might just feather their own nest. I still believe, nevertheless, that, once in a while, they will do a bit better than that, if only because they cannot help it, that is, because the constraints I originally assumed do in fact apply.

RICHARD R. NELSON, Yale University: *Exit, Voice and Loyalty* is a book that has excited many people. One reason is that there are a lot of good ideas in the book. Let me propose that the ideas are at at least three different levels. At the most general level, *Exit, Voice and Loyalty* is a complaint about the narrowness of much economic thinking regarding ways to control economic activity and a proposal that we broaden and enrich our thinking. There is a suggestion that economists can learn something by reading in other disciplines, particularly political science. At another level, *Exit, Voice and Loyalty* is a proposal for dividing up the ways in which control can be exercised into two broad categories—exit and voice. Once one recognizes diversity, one is tempted to provide some useful categories of control. Exit-and-voice is a particular categorization. At still another level there are a number of analytic propositions about the relationship of exit and voice to various variables, and the interaction between the two. For example, it is proposed that the use of exit erodes the power of voice, that in many cases the exit of demanders of high quality leads to an organization that provides shoddy quality and hence that exit imposes a negative externality on those who remain, etc.

Clearly, different people have found different aspects about the book compelling. I am among the people who have found *Exit, Voice and Loyalty* an exciting book. However, regarding the book as general complaint and proposal for richer thinking, I was persuaded of these points before. I think *Exit, Voice and Loyalty* should be regarded as one of a number of books that have been making the same general point. But its unique contribution certainly does not lie there. At the middle level of ideas, I have found the exit, voice dichotomization important and useful. Earlier, I thought that it had a very wide range of useful applicability. More recently, I have become convinced that the range of useful application is limited. For example, when I think about questions like the extent to which collective demand machinery should be substituted for private demand machinery, I have not found that the exit-voice language has helped very much. In some cases it has provided synonyms for words and concepts found elsewhere in the literature, like demand aggregation, collective choice machinery, etc., but seems to have no real advantages over other terms. And in some cases the exit, voice language seems to be simply more cumbersome than more traditional language. But for certain arenas of my interest, like my work on the organization of day care, the exit-voice language was extremely helpful, and some of Hirschman's particular propositions quite apt. It is in arenas like this that I think *Exit, Voice and Loyalty* has made a unique contribution.

I make these remarks neither to extol nor to diminish the book. Rather, my intent is to focus attention on where I think the real contribution of the book lies.

The Williamson paper develops some themes roughly similar to those I have presented above. He too is concerned that the world view aspects of *Exit, Voice and Loyalty* may be less widely fruitful than some of the

enthusiasts seem to claim. I find that I have difficulty discussing the Williamson paper without getting drawn into a detailed discussion of the *Markets and Hierarchies* proposal along the same lines that I have sketched above regarding *Exit, Voice and Loyalty:* the book as a sweeping complaint and proposal for broader thinking, as a particular set of categories set forth as a useful general purpose organizer, and a number of particular analytic propositions that go beyond the categories aspect of the book. But I must avoid getting drawn in that direction. I only shall stress my agreement with Williamson's concluding comment—that *Markets and Hierarchies* and *Exit, Voice and Loyalty* may be complementary rather than competitive. I think that is true, and because each is most productive in different limited arenas. For my purposes it is more useful to focus on the Freeman and Young papers as examples of what two scholars have made out of *Exit, Voice and Loyalty.*

Reading Freeman's paper reinforces my impressions that the exit-voice language is useful in a number of cases, but also in many instances that the use of language is forced and other terms are better. Freeman focuses on the problem of an employer of getting information regarding the trade-offs, from the point of view of the workers, among various dimensions of the job offer. He points out, sensibly, that it might be far better to talk about these matters with worker representatives than to try to find out about them merely by looking at success or failure in hiring and holding workers. Voice here has a particular meaning, and rings right in that use, although the notion that voice is a richer communication channel than exit was not one of the central themes in Hirschman's book. Freeman stresses that union negotiations which he says are the main voice mechanism, are particularly important in deciding upon certain atmospheric or public aspects of the work condition. In a sense, the union is a mechanism whereby workers can decide on how much of various public goods (which must be paid for by lower money wages) they want. Here after the first quick point is made and Freeman gets down to discussing mechanism, I think the use of the voice language is a bit strained, and more standard terms and concepts, drawn from the literature on public goods decision making, seem more appropriate. Indeed Freeman falls back upon these. Freeman says *Exit, Voice and Loyalty* has been useful to him in enabling him to think back to an earlier, more institutional literature. But I wonder if it is the specifics of *Exit, Voice and Loyalty* that are helpful here, or whether what Freeman really is welcoming is the general climate that is conducive now to institutional exploration. Hirschman's book helped to build that climate, but so did a lot of others.

Young's paper shows the same mix of fruitful and cumbersome uses of concepts from *Exit, Voice and Loyalty.* The most fruitful use is where Young focuses on situations where exit is costly (moving) and there are opportunities to influence governmental action by greater involvement. The voice concept is a good one for connoting aggressive action from within. However, when Young begins to discuss the range of considerations bearing on whether or not one should try to establish governmental machinery to try to aggregate demands for certain kinds of goods, or to take into account diffuse externalities, I am not sure that he gains much by use of the term "voice." The problem is well-described by the more traditional language of the anatomy of market failure. The issues of the appropriate organization of governmental machinery are complex, discussed in many places in the literature, and it does not help very much to have the issue posed in terms of "voice." One problem of mechanisms that aggregate diverse interests is that the voice is babble.

In other words, jointly with me, and independently, Young has made more fruitful use of the exit-voice language. It seems to me that Hirschman's concepts, and many of his specific analytic propositions, are right on the mark regarding classes of product or service with the following attributes. Experience, rather than shopping, is where the key information comes from. The services or goods can be and often should be specially tailored to the demander by the supplier. It takes time and experience for the supplier to learn about the demander's

tastes or problems. The provider has some interest in retaining his customers. The day care study that Dennis Young and I did exploits Hirschman's ideas. I think they are applicable and underused in the study of professional services more generally, for example, in the content of the patient-doctor relationship. Of course there are many other uses in which the exit-voice language is useful. Many of these are discussed in Hirschman's book. Some new arenas where exit and voice concepts can be used fruitfully are contained in the Young and Freeman papers.

I can sum up my feelings this way. The ideas in *Exit, Voice and Loyalty* are important and useful. However, the scope and generality of the big ideas sometimes have been overstated. I believe that the power of the categories, when applied to a limited but important arena, is very considerable. And some of the detailed analytic propositions in the book are very useful in certain special cases. It is a book, very useful in a circumscribed arena, like Chamberlin's work on monopolistic competition, rather than a widely embracing world view. That is reason enough for praise.

# NEW DEVELOPMENTS IN THE THEORY OF MONOPOLISTIC COMPETITION

# A Survey of Advertising and Market Structure

By GERARD R. BUTTERS*

The economics of advertising is marked by the same emotional commitment to conflicting schools of thought that is usually associated with monetary theory and the economics of speculation. To all appearances, the choice of both the axioms used and data to be interpreted has been made in order to justify preexisting conclusions rather than to make an unbiased test between alternative theories. How else is one to explain the vehemence and fixity of many economists' views in the face of gradually accumulating, but still sparse, evidence?

The controversies endure in large part because advertising still remains outside the core of accepted theory. Neither the standard Marshallian partial equilibrium analysis nor the Arrow-Debreu general equilibrium analysis even admits the existence of advertising. Two major explanations of advertising are widely held: one considers advertising as a set of psychological ploys which induce consumers to buy products or brands that they otherwise would not buy; the other treats advertising as a provision of information which allows consumers to make more discriminating choices within the framework of a fixed set of preferences. The source of difficulty is that although these theories have opposing welfare implications, it is exceedingly difficult to devise tests which distinguish between them empirically. In the absence of a sound, agreed upon theoretical base, the field began its development by asking broad and imprecise welfare questions, such as "Is advertising good or bad?" and "Does advertising reduce competition?" by developing *ad hoc* arguments on all sides, and by amassing inconclusive empirical evidence to support alternative contentions. The result is that we have accumulated a rich fund of insights, but we are saddled with imprecise terminology and ill-posed questions which stand in the way of clear analysis. For recent surveys of the literature, the reader is referred to the books by James Ferguson and Richard Schmalensee. In the space allotted here, I can only point out one example of the imprecision of the current debate and make a few suggestions as to how research might proceed.

## I. Advertising and Profitability

A case in point is the ongoing controversy over the relationship between advertising, concentration, and profitability. A number of recent studies, most prominently those done by William Comanor and Thomas Wilson (1967, 1974), have found a significantly positive coefficient for advertising in regressions of industry rates of return on industry rates of adver-

* Assistant Professor of Economics, Princeton University. I would like to thank Michael Rothschild, Larry White, and Gregory Chow for helpful discussions.

tising expenditure (usually in the form of advertising divided by sales) and other variables. This result has been held to support the proposition that advertising raises profits through some mechanism such as by creating brand loyalty or by increasing the minimum efficient firm size. Three sets of objections have been raised frequently to this line of reasoning. The first set dismisses the positive coefficient as a statistical artifact, usually on the grounds that the accounting convention according to which advertising is considered as a current expense rather than as an investment biases the coefficient upwards. The second set asserts that the positive coefficient could equally well be due to a causal effect from profits to advertising or from concentration to both advertising and profits. The third set of objections denies the existence of substantial economies of size in advertising and the existence of absolute cost disadvantages to new entrants on account of advertising induced brand loyalties, thus striking at the theoretical justifications for the advertising-profits relationship. The jury is still out on all aspects of this debate.

A second school of thought holds that not only does advertising fail to create monopoly power, but also that it is a bulwark of competition. A recent summary of this view has been presented by Yale Brozen. He argues that advertising provides useful information about firms (their existence, size, location, etc.) and their products (their existence, prices, qualities, etc.) and therefore *reduces* entry costs and makes demand curves *more* elastic than in the absence of advertising. We are meant to conclude, it appears, that because advertising enables the condition of perfect information to be more closely approximated, the standard welfare theorems regarding perfect competition apply and we can rely on the free market to provide an optimal quantity of all goods, including advertising.[1] There are two major flaws associated with this reasoning: (1) it simply assumes that advertising acts solely to inform rather than to persuade, and (2) it does not take into account the multitude of externalities that are involved in both the production and dissemination of information.

A more fundamental objection to either the view that advertising creates profits or that it enhances competition is that the level of advertising expenditures should be considered to be an endogenous variable, jointly determined along with prices, quantities of production, and rates of return. The term "cause" in a dynamic economic system may reasonably be assigned to the equations of the system, the values of exogenous variables, or even the initial values of endogenous variables. It is not reasonable to say that current values of one endogenous variable, the level of advertising, are the cause of current values of another endogenous variable, say profits.

Putting this argument in a different way, to ask whether advertising is a cause of profits in a particular industry is to ask how profit rates would change, or would have changed, in response to a change in the advertising-sales ratio. This response would differ depending on whether the source of the change in advertising was (1) a change in the tax laws regarding advertising, (2) the imposition of a quota on firms' advertising, (3) the prohibition of advertising, (4) firms' errors in determining their advertising expenditures, (5) a change in technology affecting the advertising media, or (6) some other source. Certainly, then, any question along the lines of "Does advertising cause profits?" or "Does advertising

[1] A more fair and balanced appraisal of Brozen's position would also discuss the empirical evidence he adduces to support his defense of advertising.

enhance competition?" is ill-posed without further elaboration. In reviewing the literature, one must first check that writers' positions are, or can be, posed in an unambiguous manner.

Writers who believe that advertising is in part responsible for concentration and profitability usually have one of two barriers-to-entry models in mind. One is the economies of scale argument, that the average cost of inducing a sale through advertising expenditure, presumably holding product type and price constant, falls with increasing advertising expenditure, thus leading to absolute size barriers to entry. As a matter of semantics, it is as misleading in this case to say that "advertising causes profits" as it is to say, in an industry with substantial technical economies, that "production causes profits."

In many respects economies of size in sales promotion are the same as economies of large scale production and require similar treatment.[2] However, there are two significant ways in which advertising is different: one is that many persons have the suspicion that advertising is inherently wasteful and that therefore advertising scale economies are only private, rather than public gains; the other is that advertising costs interact with demand conditions in a way that production costs do not—one can reduce advertising costs substantially simply by lowering price. To discuss these differences, however, it is necessary to construct a theory which explains advertising in an equilibrium context involving both supply and demand conditions.

The second barrier to entry model holds that existing firms' current or past advertising somehow creates "brand loyalty" which raises the cost of new firms' entry above the costs of entry that were originally borne by the existing firms. In this case higher advertising leads to greater entry cost differentials and to greater profits for existing firms. Why, then, do not firms in all industries make optimal use of advertising, leading to a situation in which all industries have identical, or only randomly varying, levels of advertising and profits?

The usual answer is to regard advertising as a proxy for an underlying, perhaps unobservable, exogenous variable, "product differentiability." According to Comanor and Wilson (1974, p. 130), this is defined as "the product and market characteristics that permit heavy advertising expenditures to differentiate effectively the products of a firm from its rivals." They continue, "Although these product and market characteristics are not easily measured, they are typically characterized by heavy advertising expenditures." The imprecision and circularity of this definition, in addition to the lack of any operational distinction between "heavy advertising" and "product differentiability," are plain to see. On the basis of this definition, we can conclude at best that the regression results show that it is some other "product or market characteristic" which determines both advertising and profits. However, Michael Porter's recent work buttresses Comanor and Wilson's position considerably at this point by providing a concrete description of some of these market and product characteristics. (See his paper, which follows this one.)

Returning now to the alternative school of thought, a study of the price of eyeglasses by Lee Benham deserves special attention because it provides a clear interpretation of the statement that advertising enhances competition. Benham's proposition is that the *prohibition* of advertising causes prices to *rise* by preventing firms from generating enough sales to

[2] Note how neatly this sidesteps the issue of what the required treatment is!

operate at an efficient scale. He tests this proposition by comparing the prices charged for eyeglasses in different states in 1963. He finds that in those states which totally banned the advertising of eyeglasses, the average price was approximately twice as high as in those states in which no restrictions were imposed.

To conclude this section, I would like to point out that the differences between the two schools of thought are not as irreconcilable as they might first appear. There is no contradiction in believing both that advertising is crucial in facilitating firms' entry into new markets and that these entry costs might be steep enough to discourage entry. It is not inconsistent to hold that indivisibilities associated with advertising on nationwide television create market power in certain industries, but that a government policy to severely curtail advertising in these industries would cause prices to become even higher.

## II. Advertising as Information and the Equilibrium Problem

A major reason why the controversy over advertising and profits is so enduring is that there is no agreement on the more basic issue of how advertising affects sales. We can be sure that firms' advertising does increase their sales, because thousands of businesses with millions of marketing research dollars to spend could not all be wrong. Yet we have only vague and conflicting notions as to why this is true. If consumers respond positively to advertising, then what is to prevent firms from advertising inferior, inexpensive to produce goods and making large profits? If firms tend to advertise goods with poor quality, then why would consumers continue to respond to advertising? To deal with such questions one cannot separate the demand and supply sides of a market as neatly as in standard partial equilibrium analysis. The quantity demanded by consumers varies not only according to the price, quality, and advertising expenditure of firms, but also according to the nature of firms' advertising strategies, which in turn depend upon consumers' demand functions. The problems of developing an appropriate equilibrium concept to handle this interdependence of consumers' and sellers' behavioral strategies is very closely related to the problems raised by Michael Rothschild in his survey of the search literature.

An ideal theoretical structure to resolve these problems would take as given four sets of parameters, representing

(1) consumer preferences conditional on consumer information and perhaps also the amount of persuasion to which the consumer is subjected,
(2) the technology of production available to the potential producers,
(3) the technology of information transfer between consumers and producers, and perhaps also the technology of persuasion, and
(4) the regulatory framework of the industry;

from these parameters and the past history of the industry, the theory would predict such variables as concentration, prices, product quality or variety, quantities produced, profit rates, and sales promotion expenditures, including advertising. Specific models within this framework would make simplifying assumptions designed to focus attention on key trade-offs, such as the consumers' choice between searching or relying on advertising and the producers' choice between achieving sales promotion via lowered prices, improved qualities, or increased advertising.

Progress toward developing such models has recently been made in several ways. Porter (1974), by exploiting often ignored aspects of demand and information transfer, namely the distinction between convenience and shopping goods and the role

of the retailer in influencing consumers' brand choices, has greatly improved the explanatory power of the usual advertising-profitability regressions. Phillip Nelson (1970) has shown that various industry characteristics may be explained by making a distinction between "search goods," whose qualities may be determined by inspection, and "experience goods," whose qualities can only be determined after purchase. The further category of a "credence good," defined by Michael Darby and Edi Karni as one whose characteristics cannot be determined reliably even after inspection, might also be exploited to explain advertising behavior.

In a more recent paper, Nelson (1974) proposes an ingenious explanation of why national consumers respond positively to advertising of experience goods, despite the fact that such advertising often contains very little direct information. The explanation is centered around the assumption that sellers differ according to their efficiency in meeting consumers' desires. Those sellers who can produce a given product more cheaply, or who can produce a better quality product at the same cost, will seek to expand their sales to earn the maximum possible returns. They will promote their products both by offering a better buy than competitors and by advertising more heavily. As a result, the most heavily advertised brands will also tend to be the best buys, and consumers will be rational in buying products solely because they are heavily advertised. A thesis proposal by Thore Johnson at Carnegie-Mellon verifies that this argument is indeed consistent by framing it in a rigorous mathematical model.

As a final example, I will outline a model of my own which focuses upon the role of advertising in informing consumers about the existence of individual producers and the prices of their products. The following assumptions are made:

(1) All consumers are alike in that they have a perfectly inelastic demand for exactly one unit of the product.

(2a) In order to buy the good, each consumer must learn the "address" of at least one seller. Each seller may advertise this information, along with the price of his product, free of charge to consumers, but at a cost to himself of $b$ units per consumer informed. However, advertising messages cannot be directed to particular consumers; the distribution of ads is assumed to be completely random.

(2b) Consumers may also receive price and address information by direct search at a cost of $c$ units per search.

(3) The product sold is homogeneous and all of a large number of potential sellers have the same constant returns to scale production function.

(4) The government ensures that firms fulfill their advertised promises to deliver the product at their advertised price.

In this market sellers' policies may be summarized by their advertising strategies; namely, their choice of how many ads to send out at which price (or prices). Consumers' strategies may be characterized by the rule according to which they decide to terminate search and purchase at the lowest price then known to them. An equilibrium is defined as a set of sellers' and consumers' strategies such that no seller can increase his expected profits and no consumer can reduce his expected costs (the purchase price plus the costs of search), taking all other agents' strategies as given.

From these assumptions one can determine uniquely the equilibrium number of ads sent out at each price and the number of sales at each price. One can perform comparative statics exercises that show,

for example, that an increase in the cost of advertising, $b$, reduces the equilibrium number of ads sent and increases both the mean and the variance of the sales price distribution. Furthermore, one can determine whether or not the market is efficient in the sense that in equilibrium the process by which consumers and sellers are brought together is conducted at the lowest possible cost. The answer is no; in all cases the amount of advertising is excessive relative to the amount of search. After a slight modification of the assumptions is made, one can ask what would happen if advertising were taxed or searching were subsidized. It turns out that if we ignore the cost of governmental intervention, a low tax on advertising (or subsidy of search) increases efficiency, but a large one totally destroys the market.

This model illustrates the wide range of questions that can be posed clearly and answered within the context of a precise mathematical model. The real trick, of course, is to formulate similar models which parallel the "real world" closely enough that we are interested in the answers. My own model is deficient in this respect, especially in that it ignores product heterogeneity; but I hope to have demonstrated at least that the search for more rigorous and precise models of advertising behavior is a desirable complement to the traditional activities of running exploratory regressions and developing verbal analyses to interpret them.

REFERENCES

**L. Benham**, "The Effect of Advertising on the Price of Eyeglasses," *J. Law Econ.*, Oct. 1972, *15*, 337–52.

**G. Butters**, "Equilibrium Distributions of Sales and Advertising Prices," *Rev. Econ. Stud.*, forthcoming.

**Y. Brozen**, "Entry Barriers: Advertising and Product Differentiation," in H. Goldschmid, et al. (eds.), *Industrial Concentration: the New Learning*, Boston 1974.

**W. Comanor and T. Wilson**, "Advertising, Market Structure, and Performance," *Rev. Econ. Statist.*, Nov. 1967, *49*, 423–40.

——, *Advertising and Market Power*, Cambridge 1974.

**M. Darby and E. Karni**, "Free Competition and the Optimal Amount of Fraud," *J. Law Econ.*, Apr. 1973, *16*, 67–88.

**J. Ferguson**, *Advertising and Competition: Theory, Measurement, Fact*, Cambridge 1974.

**T. Johnson**, "Market Equilibrium Analysis of Advertising," unpublished pap., Carnegie-Mellon 1975.

**P. Nelson**, "Information and Consumer Behavior," *J. Polit. Econ.*, Mar.-Apr. 1970, *78*, 311–29.

——, "Advertising as Information," *J. Polit. Econ.*, July-Aug. 1974, *82*, 729–54.

**M. Porter**, "Consumer Behavior, Retailer Power, and Performance in Consumer Goods Industries," *Rev. Econ. Statist.*, Nov. 1974, *56*, 419–36.

——, *Interbrand Choice, Strategy, and Bilateral Power*, forthcoming 1976.

**R. Schmalensee**, *The Economics of Advertising*, Amsterdam 1972.

# Interbrand Choice, Media Mix and Market Performance

*By* MICHAEL E. PORTER*

We have seen considerable attention paid in recent years to building information into microeconomic theory. George J. Stigler's pioneering model of the buyer's optimal investment in information held promise for conceptual and empirical investigation of the effect of information costs on market outcomes. However, developments since Stigler have largely focused on the normative properties of equilibrium in markets where information is costly. While this is important work, I will argue here that information costs can explain differences among product markets in the observed behavior of buyers and sellers and their allocative consequences, and that these market outcomes have important implications for policy. This paper will outline a framework for modeling the cross-product variation in the observed pattern and level of product information flow. One implication of differing information equilibria across product markets is that the mix of advertising media utilized by sellers will vary from industry to industry. I will argue that this variation has important implications for market performance, and will test that hypothesis empirically here.

## I. Information and Interbrand Choice

A central result of costly information is that buyers choosing among the available brands of a product (*interbrand choice*) must invest in information gathering. Interbrand choice in product markets and the associated equilibrium of information are developed in detail in a forthcoming book (Porter 1976) and can only be sketched here. Buyers facing the choice among brands have numerous sources of information available: their own experience with the competing brands, advice from friends, salesmen, technical literature such as *Consumer Reports*, advertising messages in the various media, etc. Buyers gather information to the extent that they expect it to increase their utility, net of its cost, by directing them to the brand of a product whose attributes best meet their needs. I assume a given array of brands. (The seller can influence the buyer's information gathering process by innovation which alters this array. See Porter 1976.) The cost of information is measured in utility terms, and includes not only the time and dollar outlays to secure it but also the subjective gain or loss from the act of consuming it (compare watching a television commercial and reading a technical publication). Subject to the existence of systematic differences in the expected variance of quality among brands across products, the buyer will be willing to spend more on information gathering for a product which represents an important part of his income, or for which the perceived benefits of an informed choice are otherwise high.

Across products, the overall utility the buyer derives from the product, the set of product attributes he values and his utility ranking across these product attributes will, in general, vary. Information sources inform the buyer about attributes (price, style, reliability, etc.) rather than

* Graduate School of Business Administration, Harvard University.

about whether the product is good or bad overall. Because each information source informs the buyer about a different set of product attributes, he will divide his outlay on information differently for each product depending on how he values each set of attributes. In addition, each information source involves different utility costs and each source provides the buyer with information of differing perceived reliability and usefulness which causes him to apply different weights to it in his decision. For example, information from past experience is costless but may or may not be perceived as reliable by the buyer. Information from retailers' salesmen is more costly to gather than, say, advertising that comes embedded in entertainment or news. But a salesman's presentation is considerably more versatile in addressing the particular buyer's information needs than a standardized advertising message, as well as more expert and less biased toward the manufacturer since the manufacturer rarely controls the retailer directly. From these considerations the buyer simultaneously determines his total outlay on information and the portfolio of information sources he consults, both of which vary in general from product to product (as well as from buyer to buyer). The amount of advertising, for example, varies from product to product not only because of varying efficiency for supplying messages to buyers (Lester Telser 1964, P. K. Else 1966) but also because buyers' demands vary depending on advertising's perceived cost/benefit ratio relative to other information sources.[1]

The manufacturer of a product must decide his optimal outlay on the sources of buyer information that he controls or influences. He sets advertising levels on the various media directly. He influences the information the retailer provides about his product indirectly through his provision for the gross margin the retailer receives, through sales efforts directed at the retailer using his manufacturer's sales force, and through other services he provides the retailer such as cooperative advertising, financing, etc. The manufacturer's expenditure on each advertising medium and on influencing the retailer will depend in part on the buyer's information calculus, or how responsive the buyer is to messages from the various sources. It will also depend on the cost of placing messages before potential buyers via each information channel (the technology of converting dollars to messages placed before potential buyers of the product). If the product is sold to a small proportion of households, for example, it may be less efficient to utilize media that reach all households than more specialized media that cost more per message but are subject to less leakage.

From the responsiveness of buyers to each information source (the net revenue to the seller as a function of the number of messages received from that source) and the dollar cost of placing (or influencing) messages before the buyer from each source, the manufacturer calculates his optimal portfolio of sales promotion expenditures. The seller's optimal portfolio is affected by the presence of existing and potential competitors. The set of sales promotion portfolios chosen by competing sellers results from a market equilibration in which competing advertisers tend to drive up the cost and depress the effectiveness of any one rival's messages. Thus the observed levels of outlays on information transfer in a market will depend on three equilibrating processes: the buyer's information equilibrium, the individual seller's equilibrium and the market equilibrium reflecting the interactions of seller outlays

[1] Phillip Nelson's (1970, 1974) approach, though important in its focus on the buyer's information problem, also suffers from a lack of completeness. See Porter (1976).

on each other. Previous research has by and large focused on one of those at a time.

Such a process will, in general, cause the portfolio of information channels utilized by sellers to vary among products. It will also lead to market power at the retail stage, as was shown in an earlier article (Porter 1974). For some products the information supplied by retailers is more influential in determining brand choice than the information supplied directly by the seller (typically expensive, infrequently purchased products). The retailer's influence on the buyer's choice thus should yield him the ability to extract concessions from manufacturers, and this bargaining power is evident among the statistical determinants of the profits of consumer goods manufacturers. The interbrand choice process should also affect the process by which product innovations are diffused, because buyer's varying propensities to gather information influence the channel by which new or improved products can be promoted.

More broadly, information costs and the resulting interbrand choice process impel sellers to perform two complementary functions: producing (including innovation) their product and informing buyers about it. Imperfections in the market for "informing" may distort outcomes in the market for "producing." The firm most efficient at manufacturing or innovating may not be the most efficient in informing sellers about its product. In addition, we cannot map from the observed product and technology space to buyers' utility space because buyers will differ, in general, in their degree of informedness. This latter observation complicates the empirical analysis of product variety.

The process of interbrand choice leads to market power among sellers to the extent that sellers' information lacks perfect substitutes (because sellers can supply it more cheaply than alternative sources) and that select sellers have superior access to the channels of information than their existing or potential rivals due to imperfections in its supply. Some have argued that advertising and other information provided by sellers improves buyers' choices and hence improves market performance rather than detracts from it. Where providing information yields market power to sellers, the welfare benefits of better informed choice must be weighed against the welfare costs of the market power.

## II. Advertising Media Mix and Market Performance

Included in the array of possible information sources available to buyers are the various types of advertising media: network television, spot television, magazines, newspapers, etc. Since these media have differing utility costs to the buyer, provide information about differing attributes, and differ across products in their cost efficiency of placing messages before potential buyers, a consequence of interbrand choice and market information equilibrium should be that the media mix (or vector of seller outlays on each advertising medium) will vary from product to product. Such variation in practice is apparent to even the most casual observer of consumer goods industries. Furthermore, the variation in media mix is in large part exogenous from the seller's viewpoint, depending on characteristics of the product, buyer population and the technology of the media which are beyond the seller's control.

Is advertising fungible, or does the mix of advertising media occurring in a market have implications for market performance? One often sees the argument that printed advertising (e.g., magazines and especially newspapers) is more "informative" and hence of greater social benefit. Indeed,

printed advertisements often contain prices and other relatively objective information, which is contrasted by some to "persuasive" television advertisements. While this issue is an interesting one, it resists scientific analysis. The effect of media mix on allocative performance can be analyzed objectively, however.

A number of empirical investigations of the structural sources of market power have verified the importance of the level of advertising outlays in an industry in determining the average profitability of firms in it.[2] Does the mix of media adopted by sellers, as well as level of advertising outlays they choose, influence market power? There are strong reasons to suspect that it does. Advertising's effect on market power is due not to its effect on the buyer per se but to imperfections in its supply which give some firms systematic advantages in advertising over their existing and potential competitors. These imperfections operate primarily via the mechanism of entry barrier. Advertising leads to entry barriers because its provision is subject to economies of scale, creates absolute cost advantages for going firms and enhances the capital required for entry into an industry. This theoretical link, amended somewhat from the Comanor and Wilson view, is described in Porter (1976).

But the relationship between advertising and entry barriers will vary by advertising medium. The economies of scale in advertising depend on the threshold level of advertising required of entrants to achieve parity with going firms. Consider an entrant facing a market with going firms spending $X$ percent of sales on advertising. Under the best of circumstances, the entrant must spend $X$ percent of his sales on advertising to achieve a comparable level of product acceptance. (Of course a firm may choose to spend less but such strategy will be associated with lower levels of profitability [Porter 1976].) Apart from scale economies, the entrant may suffer an absolute-cost disadvantage in that the capital cost of establishing a given market share through advertising is greater for the $n + 1$th seller than for the $n$th.

What is less often recognized, however, is that entrant's disadvantage varies markedly depending on the media used by going firms, because the size of the "effective" threshold is different for different media.[3] Suppose that the going firms compete nationally but advertise solely in local newspapers. The entrant need offer his competing product only in one or more local markets, spending on advertising only enough to achieve parity there with its national competitors. Suppose, instead, that the going firms utilize network television or magazines reaching a national audience. Here the entrant, if he is to achieve parity in advertising, must promote his product to a much larger market and spend as much on advertising as the going firms regardless of his sales volume. National magazines, and to a greater extent network television are highly indivisible media with respect to geographic market boundaries and market size, and this indivisibility increases the effective threshold level of advertising on them the entrant must vault.

The divisibility of each medium by market size also depends in part on the structure of advertising rates used in the medium as well as the variety of competitors in each media type. While media rate structures are complex, we can sketch

[2] See William Comanor and Thomas Wilson (1974) and Porter (1976). Porter (1976, Ch. 6), discusses the standard criticisms of this result.

[3] These arguments also apply to the advantages of leading firms over their smaller rivals in the industry. For an extension of the concept of entry barriers to competition *within industries*, see Porter (1976, Ch. 4), and Richard E. Caves and Porter (1975).

some suggestive implications here.[4] The seller wishing to utilize network television must beam his message over a large percentage of the network's stations, which are comparable in number for the three major networks. Network rates are also roughly comparable across networks. However, spot television rates vary markedly across TV stations as stations' audience sizes vary. Thus a potential entrant can introduce his product and initially direct his advertising outlays to small cities with lower rates, moving to the larger cities as his scale increases; this option is not available in network advertising.

Magazines vary markedly in advertising rates, also depending on circulation and on the specialization of the readership (e.g., *Time* versus *Progressive Grocer*), the more specialized and larger circulation the higher the rates. For many of the large circulation national magazines, however, zoned or regional editions allow firms to utilize the media but direct the advertisements to a less than national market at a modest cost disadvantage relative to national advertisers. Newspaper rates also vary with circulation and specialization. In addition, in any given city there are typically a number of different papers with differing levels of these indices. Like magazines, the large circulation metropolitan papers usually have zoned editions which allow the advertiser to purchase an advertisement which will be seen by only a portion of the readership of the paper. These factors reduce the effective threshold. Magazines, though typically national, can in some cases be divisible by region by the advertiser. Newspapers are local, and are divisible even below the local metropolitan area in many cases. Thus the potential entrant faces smaller drawbacks with magazines than with network television, and less still with newspaper advertising.

In addition to a medium's divisibility by market size, the effective threshold depends on the absolute cost of a single message on each medium and the economies due to increased effectiveness of repeated messages. Absolute outlays vary by medium: sponsoring a half-hour program on network television in prime time requires an outlay on the order of $50,000 per showing; a full page color advertisement in *Business Week* costs approximately $14,000; a 30-second prime time spot television in Los Angeles costs approximately $4,000; and a full page advertisement in the *Chicago Tribune* costs approximately $8,000.[5] However, it is possible to purchase newspaper and magazine advertisements in varying increments down to the small fraction of a page at linearly decreasing rates. The minimum purchase of television advertising is the 10-second announcement, which is typically 50 percent of the cost of the standard 30-second commercial. Thus while sponsors can purchase network television advertising by the individual commercial through sharing programs, the smallest individual outlay is in the thousands of dollars as opposed to the tens or hundreds for the other media. These figures themselves do not suggest high absolute costs, but if a minimum message bombardment must occur (as suggested by Comanor and Wilson), they loom more significant. The argument that repetition must occur for effectiveness of messages would seem most significant for electronic media where the message is not available for repeating readings. This would tend to reinforce the greater minimum absolute outlays that television requires.

Finally, there has been intensive controversy over whether quantity discounts

[4] All rate structure information discussed below is derived from Standard Rate and Data Service, Inc., publications.

[5] The absolute outlays for spot television, magazines and newspapers vary by publication or station.

amplify any economies of scale in the repetition of messages. (See Comanor and Wilson 1974, pp. 53–61 and James M. Ferguson 1974, p. 78.) While the presence of quantity discounts is unclear in network television, published rate schedules do suggest volume discounts in spot television and for advertising in many magazines and newspapers. Neglected in the discussion of the quantity discounts issue, however, is the presence of great economies of scale in national over local TV advertising. There appears to be a distinct cost advantage to advertising via the network versus utilizing spot advertising. Network rates range from approximately 10 to 70 percent of the sum of the individual station rates, with the discount varying by time of day and season. This means that a potential entrant cannot effectively utilize spot advertising in a limited area to counter network advertising by going firms.

Taken together, this evidence suggests that network television advertising and to a lesser extent advertising in large circulation national magazines should have the most significant impact on market power. Local newspaper advertising carries the least impact. And other media probably fall somewhere in between. In addition, the significance of magazine advertising for market power should depend on the type of magazine utilized, since the effective threshold outlays on these vary with the geographic scope and specialization of the magazine.

## III. Empirical Tests

To test for the influence of the media mix on market power, I compiled data on advertising expenditures by media in 39 of the 42 consumer goods industries used in an earlier analysis of structure-performance relations (Porter 1974).[6] The core variables used in the study are: profit after tax as a percent of stockholder's equity, eight-firm concentration ratio, minimum efficient scale as a percent of industry sales, industry growth, a dummy variable for local or regional industries, absolute capital requirements for production at minimum efficient scale, advertising as a percent of sales ($A/S$), advertising per firm for the leading firms ($A/F$) and an advertising interaction variable ($INTER$) equal to the product of $A/S$ and $A/F$. In addition, the distinction between convenience and nonconvenience goods described in Porter (1974) was utilized here. To these were added data on relative outlays on national television, spot television, magazines, national newspaper supplements and local newspapers of leading firms in each industry in the sample. Although data for some media were not available, the ones included represent a substantial majority of estimated total media outlays.[7] To that extent they accurately measure the proportional expenditures on the principal media.

The average percent of total advertising outlays of the leading firms spent on each advertising medium were computed for each industry. $A/S$ for the industry was multiplied by the percent of total advertising accounted for by outlays on each advertising medium to approximate the ratio of the advertising outlays on each medium to sales:

$NET/S$ = network television advertising/sales

$SPOT/S$ = spot television advertising/sales

$MAG/S$ = magazine advertising/sales

$LOCNEWS/S$ = local newspaper advertising/sales

[6] The sources, construction and limitations of the media data are described in a technical appendix available separately from the author.

[7] See *Advertising Age*, May 17, 1971 for an estimated breakdown of total media outlays. Radio represented only an estimated 6 percent of outlays in 1968, and network radio only .4 percent. Other omissions are discussed in the appendix.

TABLE 1—MULTIPLE REGRESSION EQUATIONS EXPLAINING PROFIT RATES INCLUDING THE MIX OF ADVERTISING EXPENDITURES BY MEDIA

| | Intercept | Eight-Firm Concentration Ratio | Minimum Efficient Scale | Growth | Regional Dummy | Absolute Capital Requirements | A/S | INTER | NET/S | NATL/S | INTER 1 | INTER 2 | $R^2$ | Corrected $R^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | All Consumer Goods Industries $n = 39$ | | | | | | | | |
| 1. | 53.57[b] | −.51709[b] | .01730[a] | .02115[a] | 17.250 | .000761[c] | .54626[c] | | | | | | .559[c] | .447 |
| | (1.965) | (1.735) | (1.540) | (1.602) | (.938) | (2.942) | (3.713) | | | | | | | |
| 2. | 59.23[b] | −.52207[b] | .02095[b] | .01551 | 25.19[c] | .000796[c] | | | 1.013[c] | | | | .590[c] | .513 |
| | (2.223) | (1.819) | (1.934) | (1.176) | (1.402) | (3.181) | | | (4.144) | | | | | |
| 3. | 64.85[b] | −.51970[b] | .01937[b] | .01198 | 18.63 | .000867[c] | | | | .8888[c] | | | .577[c] | .498 |
| | (2.357) | (1.782) | (1.762) | (.858) | (1.033) | (3.300) | | | | (3.960) | | | | |
| 4. | 58.63[b] | −.47439[a] | .02148[b] | .02108[a] | 19.41 | .000717[c] | | | | | .007752[c] | | .564[c] | .483 |
| | (2.130) | (1.613) | (1.923) | (1.610) | (1.059) | (2.840) | | | | | (3.784) | | | |
| | | | | | | Convenience Goods Industries $n = 17$ | | | | | | | | |
| 5. | 48.39[a] | −.78747[c] | .01821[b] | .02903[b] | 13.020 | .001949[c] | .48092[c] | | | | | | .888[c] | .820 |
| | (1.567) | (2.878) | (1.754) | (2.113) | (.910) | (2.827) | (4.946) | | | | | | | |
| 6. | 57.90[b] | −.82478[c] | .02205[b] | .02230[a] | 20.706[a] | .002400[c] | | | 1.0115[c] | | | | .897[c] | .835 |
| | (1.918) | (3.135) | (2.248) | (1.611) | (1.523) | (3.563) | | | (5.255) | | | | | |
| 7. | 65.37[b] | −.84556[c] | .02257[b] | .01568 | 18.032[a] | .002358[c] | | | | .93890[c] | | | .906[c] | .849 |
| | (2.224) | (3.351) | (2.408) | (1.126) | (1.384) | (3.670) | | | | (5.576) | | | | |
| 8. | 72.43[b] | −.85684[c] | .01969[b] | .02161[a] | 13.551 | .002448[c] | | | | | .008254[c] | | .894[c] | .831 |
| | (2.273) | (3.201) | (1.969) | (1.531) | (.978) | (3.579) | | | | | (5.170) | | | |
| | | | | | | Nonconvenience Goods Industries $n = 22$ | | | | | | | | |
| 9. | 113.3[c] | −.30871 | −.02748[a] | −.00359 | | .000417[a] | | .0001458[c] | | | | | .727[c] | .641 |
| | (4.356) | (.718) | (1.597) | (.239) | | (1.669) | | (4.710) | | | | | | |
| 10. | 108.4[c] | −.0905 | −.02472[a] | −.00437 | | .000428[b] | | | | | | .000298[c] | .803[c] | .741 |
| | (5.017) | (.253) | (1.726) | (.343) | | (2.017) | | | | | | (6.069) | | |

*Note:* figures in parentheses are t values. The significance of the regression coefficients is tested using a one-tail t test and the significance of the coefficients of multiple determination is tested using the $F$ test.

[a] Indicates coefficient is significant at the 90 percent level.
[b] Indicates coefficient is significant at the 95 percent level.
[c] Indicates coefficient is significant at the 99 percent level.

From these ratios the following additional variables were constructed:

| | |
|---|---|
| *NATL/S* | = national advertising/sales, the sum of *NET/S* and *MAG/S* |
| *TV/S* | = the sum of *NET/S* and *SPOT/S* |
| *INTER 1* | = an interaction variable equal to the product of *A/S* and *NET/S* |
| *INTER 2* | = an interaction variable equal to the product of *A/F* and *NET/S* |

### IV. Results

Table 1 presents multiple regression equations incorporating the media-mix variables.[8] The results, subject to the usual deficiencies in the data, are striking. For the full sample of 39 consumer-goods industries, replacing *A/S* with *NET/S* increases the size and significance of the coefficient of advertising and the corrected $R^2$ of the equation, although advertising outlays on network television represented no more than 70 percent of total advertising. *INTER 1* produces a similar improvement, though not as striking. *NATL/S* yields results slightly inferior to *A/S*. Introducing the ratios for other media singly, additively, or in interaction form yielded insignificant (and sometimes negative) coefficients.

Consistent results are obtained when we partition the consumer-goods industries into convenience and nonconvenience goods—a distinction previously found central to understanding the influence of advertising on profitability. For convenience goods *NET/S* is more significant than *A/S* and improves $R^2$, while *NATL/S* performs even better. *INTER 1* similarly improves results over the results with *A/S*. Advertising intensities for other media alone and in combination yielded no significant variables, although the coefficient of *LOCNEWS/S* was consistently negative. The inclusion of the growth-concentration interaction described in Porter (1974) did not affect the results.

In nonconvenience goods, the respecified advertising interaction variable *INTER 2* (*A/F* * *NET/S*) produced major improvements in significance level and in corrected $R^2$. Again, other specifications of the media data were unsuccessful, though respecifying the advertising interaction variable using *NATL/S* was nearly as significant as the original advertising interaction variable (*INTER*).

Taken as a group, these results suggest that the elevation of market power due to advertising is primarily due to the size of advertising outlays on network television and magazines, especially the former. There is apparently little interaction between total outlays the firms make on advertising of all kinds and the share of advertising allocated to individual media. Other forms of advertising appear to have little significance for long-run market power, though they may well be central to maintaining market shares.[9] The requirement of matching competitor outlays on these media puts entrants at no serious disadvantage which would yield going firms market power.

### V. Conclusions

Our results, though tentative, reaffirm the importance of interbrand choice and of the technology of the various information channels utilized by sellers for under-

[8] The results for the other variables besides advertising are substantially similar to those described in Porter (1974).

[9] As one might expect, the correlation between *A/S* and *NET/S* is high (approximately .9 in convenience goods, .8 in nonconvenience goods). *MAG/S* was somewhat less correlated with *A/S* (.7 and .5 in convenience and nonconvenience goods, respectively) and *LOCNEWS/S* virtually uncorrelated with *A/S*. The fact that *NET/S*, and *NATL/S* improved the sta-

standing industry performance. Although some information is surely provided most efficiently by the seller, the marginal contribution of advertising to market power must be weighed against these benefits. Second, the equilibrium in the information markets that determines sellers' proportional outlays on the various advertising media is central to industry performance, because our results suggest that the media have asymmetrical implications for market power.[10]

We have much to learn about how information equilibria vary among product markets, and what influence is wielded by the technology of the various media for providing information to buyers. Such knowledge will not only illuminate the bases for market power but also amplify our understanding of intermarket variations in: levels of advertising by media, the structure of distribution channels and how they are located, the role of independent information sources such as *Consumer's Reports,* and the differences in such traits across countries and over time.

Such analysis is also central to policy-making. Our preliminary results suggest that social limits in advertising may be relevant to network television and to a lesser extent magazines, but not clearly to other media. The process of interbrand choice also has implications for improving the degree to which buyers make informed choices. Providing more information to consumers is not likely to improve the quality of choice without change in the parameters that determine the amount and type of information that the buyer consumes. Finally, the analysis of "consumer protection" should be related to a comprehensive model of consumers' investment in information.

tistical results allows us to reject the hypothesis that $A/S$ is the relevant variable, and that these new variables performed well statistically because of this collinearity.

[10] The occurrence of asymmetrical effects of the media on profitability argues against the view that advertising's effect on market power is due to the mismeasurement of profits in accounting data. (For a statement of this position, see Ferguson (1974). If mismeasurement was the cause of the result, it is difficult to see why outlays on different media would have asymmetrical impacts in measured profitability.

## REFERENCES


R. E. Caves and M. E. Porter, "From Entry Barriers to Mobility Barriers: Conjectural Decisions and Contrived Deterrence to New Competition," *Quart. J. Econ.*, forthcoming 1976.

W. S. Comanor and T. Wilson, *Advertising and Market Power,* Cambridge 1974.

P. K. Else, "The Incidence of Advertising in Manufacturing Industries," *Oxford Economic Papers,* Mar. 1966, *18,* 88–110.

J. M. Ferguson, *Advertising and Competition: Theory, Management, Fact,* Cambridge 1974.

P. Nelson, "Information and Consumer Behavior," *J. Polit. Econ.,* Mar.-Apr. 1970, *78,* 311–29.

———, "Advertising as Information," *J. Polit. Econ.,* July-Aug. 1974, *81,* 729–54.

M. E. Porter, "Consumer Behavior, Retailer Power and Performance in Consumer Goods Industries," *Rev. Econ. Statist.,* Nov. 1974, 419–36.

———, *Interbrand Choice, Strategy and Bilateral Market Power,* Cambridge, forthcoming 1976.

Standard Rate and Data Service, Inc., *Business Publication Rates and Data,* monthly.

———, *Network Rates and Data,* monthly.

———, *Newspaper Rates and Data,* monthly.

———, *Spot Television Rates and Data,* monthly.

G. J. Stigler, "The Economics of Information," *J. Polit. Econ.,* June 1961, *69,* 213–25.

L. Telser, "Advertising and Competition," *J. Polit. Econ.,* 1964, *63,* 2–27.

# Product Differentiation and Welfare

By MICHAEL SPENCE*

One of the functions of the market system is to select the commodities that are produced and sold. This process is variously referred to as product differentiation, product selection, and monopolistic competition. Product differentiation involves a set of real economic choices because there are increasing returns or declining average costs in the development, production, marketing and distribution activities of firms. The full range of possible products is neither feasible nor desirable in the presence of increasing returns to scale.[1] Product differentiation is also an important component of imperfectly competitive strategic interaction, both static and dynamic. Some would argue it is, in many industries, the most important part of the dynamics of competition.

My main purpose here is to discuss some of what has recently been learned about the welfare aspects of product differentiation and monopolistic competition in a market system. In order to do this, it is necessary to measure welfare, and I shall adopt the strategy of measuring welfare with the multiproduct net surplus, the difference between the benefits to consumers, gross of revenues paid, minus the costs of production.[2] A product's marginal contribution to surplus is the additional gross surplus it adds, minus the cost of producing it. That in turn is the area under the inverse demand function minus the costs of production. The total surplus, of course, is not the sum of these marginal surpluses; typically the total net surplus will be greater for products which are substitutes.

Everyone who is sensitive to the intricacies of market structure is aware that product differentiation has interesting information aspects, and that there is a collection of activities associated with marketing that bear on the demand for and sales of a product. I do not deal with these here, though they are important. But so are the limits of the price system in selecting products and product characteristics. These are the subject of this paper.

If firms are not perfect price discriminators, then the revenues of a firm selling a differentiated product are not equal to that product's contribution to gross surplus. Similarly profits are not equal to its net contribution to surplus. Revenues and profits are less than the gross and net contributions to surplus respectively. But profitability is the criterion by which products are selected or rejected in a market system. In the light of the divergence between benefits and profits, it seems reasonable to suppose that the criterion of profitability may not always lead to de-

* I am grateful to Richard Caves, Richard Nelson, and Oliver Williamson for their comments. Support for this research from the National Science Foundation and the John Innon Guggenheim Foundation is gratefully acknowledged.

[1] One can think of products being points in a continuous spectrum of attributes. The market selects a finite subset of points in the continuum, these being the products actually produced. For recent contributions to this subject, the reader can refer to Robert L. Bishop, Edward Chamberlin, Kelvin Lancaster, Bruce Owen and Spence (1975), and Spence (1975).

[2] For a discussion of the accuracy of this measure, see Robert D. Willig (1973).

sirable results. Indeed, that is the case.

This paper divides into three sections. In the first, I have attempted to summarize the forces that generate welfare problems in product selection, and the nature of the problems generated. There are several of these forces, and some of them work against each other. It is nevertheless useful to identify them and to understand the direction in which they push market outcomes. The second section deals with the magnitudes of welfare losses. In the past, some of the costs of imperfect competition have been measured by the cost of the nonmarginal cost pricing of the existing set of products. Using some numerical examples, I have tried to show that a significant fraction of the cost of imperfect competition may be due to the currently unmeasured cost of having too many, too few, or the wrong products. This analysis is far from decisive empirically, since it is based on numerical examples. It is however suggestive, that over a range of assumptions about the structure of demand and costs, product selection failures may be significant components of welfare costs.

In the third section, I have tried to outline, somewhat briefly, how the empirical study of product differentiation in an industry might proceed. The problem here is that the welfare analysis requires the estimation of the demands for products that currently do not exist. This line will be pursued more fully in subsequent papers.

The reader may find it useful to read this paper with the concrete case of television programming in mind. Pretend that pay television on the cable makes per program charges feasible.

## I. Welfare Problems in Product Selection

There are several market forces operating in the selection of products.

### *A. Profits and Net Surplus*

In the absence of price discrimination, the profits of the seller of a product will fall short of the net surplus generated by the product. Therefore, a product can have a positive potential net surplus and at the same time, be incapable of generating positive earnings. The exception to this principle is the case of constant or diminishing returns to scale in production. The fact that profits do not capture all of the net surplus is a conservative force, tending to eliminate products that should be made available. I hasten to add that this is one of several forces. It does not warrant the conclusion that there are too few products, or too little variety.

The tendency to lose desirable products is greater, the more substantial are the increasing returns or fixed costs.

### *B. Falling Average Costs*

In a monopolistically competitive market, without entry barriers beyond those imposed by the usual profitability requirement, prices are set above marginal costs, and entry occurs until profits are driven to zero, or equivalently until prices are equal to average costs. These statements technically apply to products that are symmetrically different in the way that it has become customary to assume in the theory. Since price is at average cost and above marginal cost, marginal cost is below average cost and average costs are declining at that point. It does not follow from this fact that there are too many products. It is true that there are more products than there would be if nonnegative profits were required and price were equal to marginal cost. It is also true that prices equal marginal costs in the first best optimum. But it is not true that the first best or second best outcome is achieved when profits go to zero. Profits can be above or below zero when the

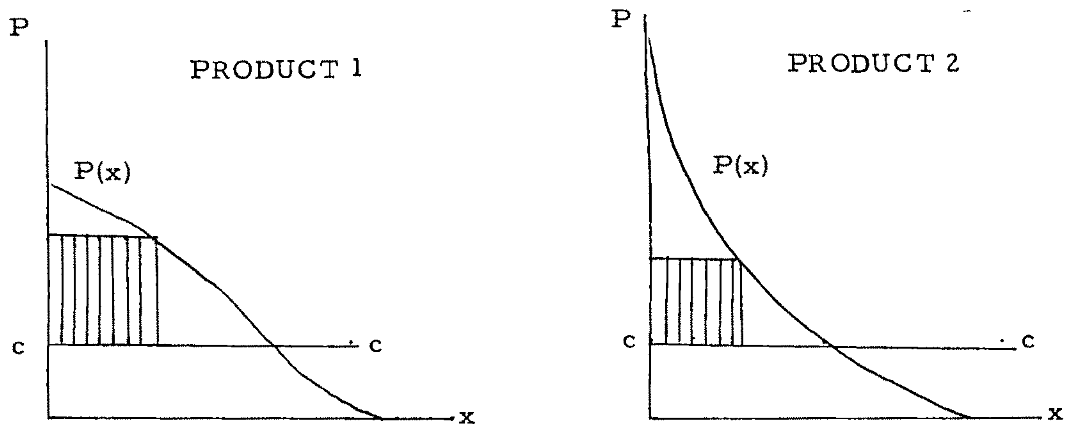



FIGURE 1.

marginal contribution of a product to the surplus falls to zero.[3]

This point is most easily seen for the case in which average costs decline indefinitely. In that case, if price equals marginal costs, profits are always negative. But it is not true in general that the optimal number of products is zero.

### *C. Revenues, Gross Surplus, and Biases Against Products*

Products are not equally subject to the conservative force discussed under (1) above. There are biases in product selection. Although revenues, in the absence of price discrimination, do not capture all of the gross surplus generated by a product, the fraction of the gross surplus that is or can be captured varies from product to product. The ratio of revenues to gross surplus depends upon the properties of the demand functions. Rather than analyze this in the abstract, let me illustrate the point with the case of constant elasticity demands. For the time being, I suppress explicit consideration of interproduct interactions. Let the price elasticity of demand for a product be $\eta$. It is easy to establish that at any price, the ratio of revenues to gross surplus (the latter being the area under the inverse demand function) is

$$\beta = 1 - 1/\eta. \tag{1}$$

Therefore, as $\eta$ rises, the fraction $\beta$ rises. The fraction of gross benefits that appear as revenues to the selling firm is larger, the higher the price elasticity of demand.

Since products vary in elasticities, it is quite possible for a product with a low price elasticity to have a higher net surplus and lower profits than a product with a high price elasticity of demand. This has two implications. The tendency to lose valuable products, discussed earlier, is stronger for low elasticity products. And second, the process of selection of products in a differentiated product industry will be biased against the lower elasticity candidates.[4]

The nature of the bias is illustrated in Figure 1. The first product has a lower

[3] The content of the so-called excess capacity theorem is therefore limited to the statement that at the equilibrium, average costs are still falling.

[4] This conclusion appears to run counter to the intuition developed in countless students of microeconomics, that low elasticities are nice for sellers because they permit higher markups. I don't wish to try to overturn this truth, but rather to point out nevertheless, that the surplus is harder to capture with the price system for low elasticity products. Let me put this more positively. I would expect the incentive for sellers to try to price discriminate would be higher for lower elasticity products.

surplus and higher profits than the second. The profits of both products in Figure 1 are positive. However, if there were a fixed component of costs, then the first product could be profitable and survive, while the second would not, even though the second one contributes more to the surplus. If the two products are substitutes, and if the market mechanism selects the more profitable, then again the first product will be taken.

If one does this analysis more rigorously, it turns out that it is *not* exactly the elasticity that matters, but rather what fraction of net potential surplus for a product is capturable by the selling firm. This depends upon both the structure of demand and of costs. But the qualitative nature of the bias is clear. If the potential consumers of a product have a highly variegated set of willingnesses to pay for it, so that there is a small group with a high willingness to pay, and then rapidly declining reservation prices after that, then the selling firm will have difficulty capturing the surplus, and the product will have difficulty being supplied under the market system. With some caution, one might refer to such products as special interest ones. They tend to be supplied by clubs or other institutional devices that do not use the pure price system.

### *D. Interactions Among Products*

Thus far, I have discussed forces that affect a single product. Let me turn now to interactions among products that are imperfect substitutes. Apart from biases in selection, interactions on the demand side determine whether there are too many or too few products, too much or too little variety. In what follows, I shall assume that products are substitutes, that they have fixed and variable costs, the latter being linear or convex, and that market equilibria are Nash equilibria in quantities. From a social point of view, the fixed component of cost can be regarded as an entry fee for having an additional product.

Perhaps the best way of approaching product interactions is to describe the circumstances that lead to too many products. They are two. The first is high own price elasticities, for these cause firms to contract output further away from the level at which price equals marginal cost. And that leaves "room" for further entry, room which would be absent with marginal cost pricing. But high own price elasticities are insufficient. The cross elasticities must also be high.

The entry of an additional product has several effects. It increases the surplus from the new product, but lowers the demand for existing products and causes them to contract output. In terms of the surplus, there are gains and losses. The gains are the profits and the consumer surplus from the new product. The losses are reductions in the profit and surplus from the existing products. When the products are close substitutes and the cross elasticities are high, the extra surplus created by the entering product is lost through contractions of existing firms. The familiar metaphor of expanding and slicing the pie is appropriate. A new product expands the pie and causes it to be sliced into more pieces. If the cross elasticities are high, the expansion is not large. But since there is a cost (the fixed cost) of adding a slice, the costs may outweigh the benefits. These forces will be illustrated in the calculations of the next section.

### *E. Half a Loaf is Better than None*

While profitability is not the correct criterion for deciding on the entry of a product, it is, for practical purposes, the only criterion we have. Nor is it likely that the information required to provide

the correct subsidies will ever be available. One can reasonably accept profitability as a constraint, and pose the problem of product selection as that of determining the right set of products subject to that constraint. The solution to the problem includes specification of not only the products, but also the prices. The prices will typically be above marginal cost, since that may be required to increase the profitability of products to permit the entry of products that are not profitable under marginal cost pricing. In short, the solution to the second best problem will include a trade-off between numbers of products on the one hand and the inefficiency due to nonmarginal cost pricing on the other.

Roughly, that is what occurs in a monopolistically competitive market. Therefore, the monopolistically competitive equilibrium has the qualitative features of the solution to the problem of maximizing the surplus subject to the condition that all products are profitable. It can happen that the equilibrium and the second best optimum coincide.[5] But in general, this coincidence is unlikely, and the best that can be said is that the equilibrium approximates the constrained optimum. The calculations that follow will give some insight into the question of how good the approximation is.

### *F. Complements*

I want to comment briefly on complementary products. Complementary products tend to be undersupplied by imperfectly competitive industries. That is to say, the prices are too high and the number of products is too small. The reason is that the tendency to hold back quantities reduces the demand for other products and makes entry more difficult, not easier, as in the case of substitutes. This means that all forces work in the same direction. The failure of firms to price at marginal cost reduces options for other firms, and entry is retarded, as well as expansion by existing firms. One would therefore expect complementary products to be supplied by multiproduct firms.

[5] See Avinash Dixit and Joseph Stiglitz (1974) and Spence (1976).

## II. Numbers and Variety of Products: Numerical Examples

In this section, I have taken some quite simple numerical examples, in which demand functions are linear and products are symmetrically different, and calculated the market equilibrium and two different optima, for a variety of the values of the parameters. The purposes of this are twofold. First, the interaction effects discussed above can be seen to work in the examples. Second, I want to show that the fraction of the total welfare loss that is attributable to the nonmarginal cost pricing of the equilibrium products, varies considerably, and is frequently less than half of the total welfare loss. The calculations also show that the equilibrium is often a reasonably good approximation to the constrained optimum, where the constraint is that profits be nonnegative.

Briefly, the basis for the calculations is the following. The quantity of the $i$th product is $x_i$. The inverse demand for the $i$th product is

$$p_i = a - 2bx_i - 2d \sum_{j \neq i} x_j. \tag{2}$$

The cost function for the $i$th firm is

$$F + cx_i. \tag{3}$$

In the calculations, $c$ is taken to be one. The remaining parameters, $a$, $b$, $d$, and $F$ vary. $F$ is the fixed cost, $d$ is the interaction effect; $a$ and $b$ are the intercept and slope of the inverse demand for each product, when there are no other prod-

TABLE 1

| Group | Case | $a$ | $b$ | $d$ | $f$ | $T_1$ | $\Delta T_2$ | $\Delta T_3$ | $N_1$ | $N_2$ | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| I | 1 | 10 | 1 | .5 | 1 | 28.8 | 2.8 | 1.9 | 5.4 | 9.7 | 1.4 | .71 | .84 |
| | 2 | | | | 2 | 24.5 | 3.5 | 1.8 | 3.5 | 6 | 2 | 1 | 1.29 |
| | 3 | | | | 4 | 19 | 4.3 | 1.2 | 2.2 | 3.36 | 2.82 | 1.41 | 2.06 |
| | 4 | | | | 6 | 15.3 | 4.8 | .6 | 1.59 | 2.19 | 3.46 | 1.73 | 2.81 |
| | 5 | | | | 8 | 12.5 | 5 | .2 | 1.25 | 1.5 | 4 | 2 | 3.6 |
| | 6 | | | | 10 | 10.3 | 5.1 | 0 | 1.01 | 1.02 | 4.47 | 2.24 | 4.4 |
| II | 7 | 10 | 2 | 1.5 | 2 | 8.2 | 2.6 | 1.4 | 1.17 | 2.58 | 2 | .71 | 1.03 |
| | 8 | | | | 4 | 6.3 | 3 | .87 | .73 | 1.3 | 2.8 | 1 | 1.8 |
| | 9 | | | | 6 | 5.1 | 3.2 | .33 | .53 | .78 | 3.46 | 1.22 | 2.7 |
| III | 10 | 10 | 1 | .1 | 2 | 99.7 | 8.7 | 1.7 | 21 | 26 | 1.49 | 1 | 1.3 |
| | 11 | | | | 5 | 56.6 | 13 | .1 | 10 | 9.4 | 2.36 | 1.58 | 2.4 |
| | 12 | | | | 7 | 39.6 | 14.8 | 2.2 | 7.1 | 5.1 | 2.79 | 1.87 | 3.2 |
| | 13 | | | | 9 | 27.3 | 16.3 | 7.3 | 5.2 | 2.2 | 3.16 | 2.5 | 4.0 |
| | 14 | | | | 10 | 22.5 | 16.8 | 11.3 | 4.5 | 1.1 | 3.3 | 2.12 | 4.4 |
| IV | 15 | 10 | .7 | .5 | 2 | 29.9 | 4.8 | 3.5 | 2.4 | 5.7 | 3.2 | 1.2 | 1.5 |
| | 16 | | | | 4 | 26 | 6.2 | 3.8 | 1.6 | 3.5 | 4.5 | 1.7 | 2.3 |
| | 17 | | | | 6 | 23.2 | 7.1 | 3.5 | 1.2 | 2.5 | 5.4 | 2.0 | 3.1 |
| | 18 | | | | 8 | 20.9 | 7.7 | 3.0 | 1 | 2 | 6.3 | 2.4 | 3.8 |
| V | 19 | 10 | .3 | .2 | 2 | 82.1 | 7.6 | 6.2 | 4.5 | 10.3 | 4.5 | 1.8 | 2.1 |
| | 20 | | | | 6 | 69 | 12 | 7 | 2.4 | 5.1 | 7.7 | 3.2 | 4 |
| | 21 | | | | 10 | 61 | 14 | 7.5 | 1.75 | 3.5 | 10 | 4.1 | 5.6 |
| | 22 | | | | 15 | 53.6 | 16.1 | 6.8 | 1.3 | 2.5 | 12.2 | 5 | 7.5 |
| VI | 23 | 10 | .3 | .05 | 10 | 170 | 23 | 2 | 9.2 | 11 | 6.3 | 4.1 | 5.6 |
| | 24 | | | | 20 | 102 | 29 | 0 | 5.1 | 4.6 | 8.9 | 5.8 | 9.4 |
| | 25 | | | | 30 | 62 | 32.9 | 10 | 3.2 | 1.72 | 11 | 7.1 | 13.3 |

ucts.[6] Note that because of the symmetry of the example, an equilibrium is described by $x$, the output per firm and $n$, the number of firms.

In Table 1, $T$ is the surplus, $N$ is the number of firms, and $X$ is the output per firm. The table shows the equilibrium, the optimum, and a constrained optimum. The constrained optimum takes the equilibrium number of firms as given and prices the goods at marginal costs. The difference between the surplus here and the optimal surplus is the loss due to the nonmarginal cost pricing of existing products. The remainder of the welfare loss is due to product selection; in this context, having the wrong number of products.

A word about the table is in order. There are six groups of calculations, distinguished by own and cross elasticities. Within each group, fixed costs increase. Thus in group I, the own slope is -1, the cross partial -.5, and the fixed costs increase from 1 to 10. The numbers reported for each case are as follows. The subscript refers to the following cases:

1 — the optimum
2 — the market equilibrium
3 — the equilibrium number of firms with marginal cost pricing.

[6] The total surplus here is

$$T = \sum_{i=1}^{n} \int_0^{x_i} p_i(x, \ldots, x_{i-1}, s_i, 0, \ldots, 0) ds_i$$
$$= \sum_{i=1}^{n} \left( a - 2d \sum_{j=1}^{i-1} x_j \right) x_i - b x_i^2.$$

When $x_i = x$ for all $i$, the surplus is

$$T = n[ax - bx^2 - d(n-1)x^2].$$

The variables are

$X =$ output per firm
$N =$ the number of firms
$T =$ surplus
$\Delta T =$ the difference between the surplus for that case and the surplus at the optimum.

Notice that $N_2 = N_3$ so only $N_2$ is reported. $\Delta T_2$ is the welfare loss in the market equilibrium. $\Delta T_3$ is the welfare loss if the equilibrium goods were priced at marginal cost.

The calculations reported in the table illustrate a number of points. First the equilibrium number of products can be above or below the optimum. Cases of too few products tend to occur when cross elasticities are low relative to own elasticity, and fixed costs are high (see Groups III and IV). I should say here that linear demands have a strong tendency to produce too many products because the profit maximizing quantity is half of that when price equals marginal cost.

The welfare loss from the nonmarginal cost pricing of existing products ($\Delta T_3$) varies considerably. Often it is less than half the total welfare loss ($\Delta T_2$). Moreover, the cases where a substantial fraction of the welfare cost is associated with the wrong number of products vary. When cross elasticities are high (Groups I, II and V), they occur when fixed costs are low. When cross elasticities are low, they occur when fixed costs are high. These are also the situations where the optimal and equilibrium number of firms differ by a large percentage.

With nonlinear demands, similar effects hold; I do not have the space to report on them. The excessively strong tendency to get too many products can be avoided with other kinds of demands.

## III. Hypothetical Products and Product Characteristics

A natural starting point for theoretical welfare analysis of monopolistic competition is the multiproduct structure of demand. For welfare purposes, preferences for actual and potential products are the proximate determinants of welfare problems. What the underlying attributes are, that differentiate the products, are not by themselves directly relevant. In terms of location problems, locations by themselves have no welfare consequences until the distribution of purchasers is specified. The same applies to products located in higher dimensional spaces of attributes.

When one turns to the task of implementing the welfare analysis of product differentiation empirically, the importance of the attribute space in which products are located, is elevated considerably. The empirical study of product differentiation requires the estimation of the demand for hypothetical or potential, as well as actual, products. This is a difficult problem, to say the least. The most natural way to approach it seems to me to attempt to estimate the structure of preferences for attributes rather than products. The rationale for taking this approach is that if the distribution of preferences for attributes in the consuming population can be estimated, then the demand for any set of products (not just the existing set) can be computed from that distribution. Attributes then, provide one way of dealing with the demand for potential products.

There is a considerable literature on this general subject, much of it centering on probabilistic discrete choices in transportation demand.

The goal of the theory is to derive functional forms for the demands for products as a function of their attributes. Attributes include price, physical and perform-

ance characteristics, aspects of service and so on.[7]

Using the attributes approach, one can in principle, compute the demand for any set of products, not just the existing set. In that sense, the conceptual apparatus permits one to compute the demand for hypothetical products, and therefore with a certain amount of cost data to draw welfare conclusions about the product differentiation aspects of industry performance. In future work, I hope to report in greater detail on the attributes approach and its application to industry and product data.

[7] For a discussion of probabilistic discrete choice models, see McFadden (1973).

## REFERENCES

**R. L. Bishop**, "Monopolistic Competition and Welfare Economics," in R. Kuenne (ed.) *Monopolistic Competition Theory*, New York 1967.

**E. Chamberlin**, "The Product as and Economic Variable," *Quart. J. Econ.*, 1953.

**A. Dixit and J. Stiglitz**, "Monopolistic Competition and Optimum Product Diversity," Institute of Mathematical Studies in the Social Sciences, technical report no. 153, Stanford 1974.

**K. Lancaster**, "Socially Optimal Product Differentiation," *Amer. Econ. Rev.*, Sept. 1975, *65*, 567–85.

**D. McFadden**, "Conditional Logit Analysis of Qualitative Choice Behavior," in P. Zarembka (ed.) *Frontier of Econometrics*, New York 1973.

**B. Owen and M. Spence**, "Television Programming, Monopolistic Competition and Welfare," *Quart. J. Econ.*, 1975.

**M. Spence**, "Monopoly, Quality and Regulation," *Bell J. Econ.*, Fall 1975.

———, "Product Selection, Fixed Costs, and Monopolistic Competition," *Rev. Econ. Studies*, forthcoming 1976.

**R. Willig**, "Welfare Analysis of Policies Affecting Prices and Products," research memo no. 153, Center for Research in Economic Growth, Stanford Univ. 1973.

NEW YORK: RIPPLE OR TIDAL WAVE?

# The New York City Fiscal Crisis: What Happened and What is to be Done?

*By* EDWARD M. GRAMLICH*

The New York City fiscal crisis as it was played out in the nation's newspapers this year had all the elements of a first class drama. There was first of all the tension—would the city make it through its periodic financial hurdles, would the Ford Administration blink, what would happen if the city defaulted? Then there were the accusations—was it the fault of Wagner, Lindsay, Beame, Rockefeller, Ford, the unions, or economic and social forces beyond the city's control? Then the controversy—the issue seemed ideally suited to split deficit spenders from budget balancers, soft-headed liberals from hard-headed accountants, eastern establishment intellectuals from the silent majority. Finally, though it did not capture as much press coverage, the crisis also graphically illustrated several basic issues in the economics of federalism that are now creeping into public finance textbooks—the proper role of local and national governments in stabilizing the economy and redistributing income, whether the federal or the state government has an obligation to protect the financial integrity of local governments, whether public expenditures can be effectively controlled in the short run. This paper discusses the city's fiscal plight in the context of all of these issues.

* The Brookings Institution and Cornell University. I have benefited from discussions with Robert Reischauer and Harvey Galper and the assistance of Jack Whiting.

## I. How Big are the Deficits?

Local governments borrow for three reasons, two of which are generally acceptable to bond holders and one dangerous. The first acceptable reason is to finance long-term capital investment projects—if a locality is building a school that will last for many years, it does not have to pay the entire bill in any one year but can spread the cost over the lifetime of the school by floating long-term debt. The second acceptable reason is to smooth out seasonal fluctuations in revenues and expenditures by short-term borrowing. The final reason, which generally is not acceptable either by a city's laws or in the eyes of bondholders, is to cover a current account deficit.

The first fact to recognize about the New York City fiscal plight is that unfortunately the city has borrowed to cover current account deficits continuously ever since fiscal 1960–61. How this was allowed to happen is still something of a mystery, but the relevant data, from the official U.S. Census figures, are given in Table 1. The first column lists the gross revenue of the city—taxes, charges and fees, grants from the federal government and New York State, and revenue from the water and transit authorities (which have been combined with the general government in this analysis). The second column lists the current general government expendi-

TABLE 1—NEW YORK CITY BUDGET DATA, FISCAL 1960–1974
(Millions of current dollars)

| Date | Revenue | Current Expenditures Plus Debt Retirement | Current Account Surplus | Capital Expenditures | Net Borrowing |
|---|---|---|---|---|---|
| 1960 | 2,769.6 | 2,726.5 | 43.1 | 528.3 | 485.2 |
| 1961 | 2,901.0 | 2,948.2 | − 47.2 | 542.0 | 589.2 |
| 1962 | 3,119.5 | 3,170.0 | − 50.5 | 582.0 | 632.5 |
| 1963 | 3,408.4 | 3,459.0 | − 50.6 | 667.4 | 718.0 |
| 1964 | 3,688.8 | 3,788.5 | − 99.7 | 657.6 | 757.3 |
| 1965 | 3,961.1 | 4,015.4 | − 54.3 | 657.7 | 712.0 |
| 1966 | 4,367.5 | 4,537.0 | −169.5 | 561.3 | 730.8 |
| 1967 | 5,174.8 | 5,176.0 | − 1.2 | 554.8 | 556.0 |
| 1968 | 6,085.8 | 6,144.1 | − 58.3 | 658.0 | 716.3 |
| 1969 | 6,864.7 | 6,945.6 | − 80.9 | 698.3 | 779.2 |
| 1970 | 7,233.9 | 7,775.4 | −541.5 | 797.0 | 1,338.5 |
| 1971 | 8,274.8 | 9,053.9 | −779.1 | 1,135.1 | 1,914.2 |
| 1972 | 9,501.5 | 10,119.6 | −618.1 | 1,192.9 | 1,811.0 |
| 1973 | 10,774.9 | 10,807.2 | − 32.3 | 1,371.5 | 1,403.8 |
| 1974 | 11,291.5 | 11,779.1 | −487.6 | 1,709.6 | 2,197.2 |

*Source:* U.S. Bureau of the Census, *City Government Finances;* various issues.

tures for all functions, including water, transit, contributions to the pension retirement system, and expenditures to retire maturing debt (which can be considered as the measure of capital consumption expenditures when investment is financed by long-term debt). The deficits are in the third column, the level of capital expenditures in the fourth, and net borrowing in the fifth. The large current account borrowing, particularly in recent years, was not backed up by capital formation; as a consequence it had to be done in the short-term market, and the city now has both extraordinary levels of outstanding short-term debt and extraordinarily high "uncontrollable" expenses simply to roll over this short-term debt every year. In 1973–74, for example, the outstanding short term debt was $485 per capita, much higher than any other large city, and it would now claim 32 percent of current expenditures—almost the entire controllable portion of the expenditure budget—to pay it off if new lenders cannot be found.[1]

While these are certainly not reassuring figures, they are based on the latest available Census information and understate the magnitude of the current fiscal problem in three important ways. The first is that budget data are only available up to fiscal 1973–74, which is before the onset of the current recession that cost the city dearly. Had the city's unemployment rate in 1973–74 been what it is today (11 percent instead of 6.5 percent), the deficit would have been larger by about $500 million.[2] The second is that the Census

[1] Extensive data on levels of short-term debt per capita can be found in Data Resources, Inc., "New York City Default: Some Economic Implications." Only two cities, Rochester and Yonkers, have higher levels of short-term debt per capita. Rochester has a high credit rating and seems in fairly sound position; Yonkers has also been flirting with default all year.

[2] This estimate is derived as follows. Nationally incomes are reduced by 2.8 percent for every 1 percent increase in the unemployment rate. Were this to be the case in New York City, 1973–74 incomes would have been lower by 12.6 percent in this hypothetical recession. If income and sales tax revenues decline proportionately and property tax revenues and charges slightly, the revenue loss from lower income would have been about 7.5 percent, or $400 million. On the expenditure side, the increase in unemployment translates into a 130 thousand loss of private jobs. Nationally it has been estimated that one private job loss raises the poverty population by one person—were

basically takes the city's word on the breakdown between current and capital expenditures, and it has been admitted by city officials that many current expenditures were inappropriately classified as capital items. Were the more realistic Municipal Assistance Corporation (*MAC*) figures used, roughly $700 million would come out of the capital account and go into current expenditures, raising the current deficit by the same amount. The third is that it is generally conceded that the retirement pension funds, based actuarially on mortality tables from many years ago, are underfunded, meaning that present general government contributions do not cover the annual increase in liabilities of the funds. There is no good estimate of the extent of this bias, but it is perhaps not unreasonable to say that the true contributions should be larger by 25 percent, or $200 million. Were all of these changes made, the measured current account deficit of $488 million in 1973–74—already a very large number—might have been realistically estimated to increase by about $1,400 million, giving a predicted "low employment" current account deficit of nearly $2 billion, or about seventeen percent of revenues. The anticipated deficit for fiscal 1975–76 is somewhat smaller than this—approximately $800 million for the New York City version of the current account deficit, another $600 million of operating expenditures in the capital budget, and a still unknown shortfall in pension contributions—in part because of the cutbacks mandated by the New York State Financial Emergency Control Board.[3] Even with these cutbacks, however, the general picture is one of large and growing current account deficits, persisting for long periods of time, and bound to frighten prospective lenders and threaten default sooner or later.

### II. The Composition and Growth of the Deficits

The hypotheses that have been advanced to explain these deficits can be divided loosely into two groups. One type of explanation focuses on the power of the public employee unions in New York City, the bargaining concessions that have been won over the years, and the sheer size and ability to control votes of public employee unions.[4] A second focuses on the fact that the city finances an ambitious social welfare program on its own—it funds expensive welfare, higher education, public hospital, and public housing programs, along with a persistently large transit deficit and generous contributions to employee pension funds. In other large cities these programs are either not undertaken at all by the public sector, completely financed by grants or user charges, financed by the tax revenues of independent special districts, done by an overlapping county government or done by the state. Though all programs might be desirable at the national level, and in fact many argue that the city is doing a job made necessary by the lack of adequate national or state policies in these areas, it remains an inescapable fact of federalism that if localities try to redistribute income to a much greater extent than the other locali-

this true in New York, the private job loss would raise the welfare population and net expenditures by about 13.5 percent, or $100 million.

[3] See the Joint Economic Committee Report, *New York City's Financial Crisis*, November 3, 1975, p. 30 ff.

[4] A long catalogue of the bargaining concessions is given by Raymond D. Horton. An indication of the sheer size effect is given by the fact that there are now about 450,000 full and part-time city government employees in New York City. If each was married, lived in the city, and had one close friend or relative who would vote alike on city issues, conceivably 1,350,000 votes, 30 percent of the entire voting age population and roughly half of the probable number of voters, could be marshalled in favor of making some strategic concession to, or dealing leniently with, unions.

TABLE 2—COMPOSITION OF NEW YORK CITY BUDGET DEFICIT, FISCAL 1974
(Millions of current dollars)

| | | | |
|---|---|---|---|
| Revenues | 11,291.5 | Current expenditures | 11,779.1 |
| Grants | 5,076.3 | Normal functions | 3,769.5 |
| Normal functions plus untied | 1,442.2 | Schools | 1,726.3 |
| Schools | 872.4 | Marginal functions | 6,283.3 |
| Marginal functions | 2,761.7 | Welfare | 2,587.4 |
| Welfare | 2,393.1 | Higher education | 490.1 |
| Higher education | 196.0 | Transit | 989.7 |
| Public hospitals | 135.5 | Public hospitals | 1,088.1 |
| Public housing | 37.1 | Public housing | 294.3 |
| Own revenues | 6,215.2 | Pension contributions | 833.7 |
| Taxes plus normal charges | 5,186.0 | | |
| Marginal functions | 1,029.2 | | |
| Higher education | 72.0 | | |
| Transit | 629.5 | | |
| Public hospitals | 147.3 | | |
| Public housing | 180.4 | | |
| Current account deficit | 487.6 | | |
| Normal revenues | 6,628.2 | Normal expenditures | 3,769.5 |
| School grants | 872.4 | School expenditures | 1,726.3 |
| Current account deficit | 487.6 | Marginal functions deficit | 2,492.4 |
| | | Welfare | 194.3 |
| | | Higher education | 222.1 |
| | | Transit | 360.2 |
| | | Public hospitals | 805.3 |
| | | Public housing | 76.8 |
| | | Pension contributions | 833.7 |

*Source:* U.S. Bureau of the Census, *City Government Finances*, 1973–74.

ties, or if they run abnormally large deficits and incur subsequent high interest and debt retirement costs, there is a good risk that the taxpaying population will simply pick up stakes and leave the locality in a fiscal situation that much more precarious.

These competing hypotheses can be tested very loosely by examining the city's budget in more detail. Table 2 decomposes expenditures into a normal component—expenditures for functions normally undertaken by city governments—school expenditures—which are usually financed out of the budgets of independent school districts in other large cities—and the net deficit of each of these six "marginal functions"—marginal in the sense that they do not normally drain revenue from the city government. The total deficit on these marginal functions came to the enormous total of $2,492 million in 1973–74, almost five times the measured current account deficit.[5] It is of course true that often the deficits in the accounts of these marginal programs are related to union pressure, particularly in the case of pensions but also in the case of transit, hospitals, and higher education, but as an overall impression it does appear that the city's problems stem

[5] The Census gives total public welfare expenditures of $2,587 million and grants of $2,393 million in 1973–74, leading to a net deficit on the welfare account of only $194 million. Adjusting this figure to a 1975–76 basis would perhaps add another $100 million. This Census estimate of the drain due to welfare is just one-third of that given by a recent Congressional Budget Office report on New York City, which relied both on program information and on information from *MAC*. The difference is apparently caused by the fact that Census must be putting some medicaid expenditures into the public hospital category, but not medicaid grants. Correcting this problem with the Census data would alter the composition of the marginal functions deficit between welfare and public hospitals, but not the overall total. See Congressional Budget Office.

TABLE 3—THE GROWTH OF THE NEW YORK CITY BUDGET, FISCAL 1960–74
(Millions of current dollars)

| Fiscal Year | Normal Revenue plus School Grants | Normal Expenditures | School Expenditures | Marginal Functions Deficit | Current Account Surplus |
|---|---|---|---|---|---|
| 1960 | 2,154.5 | 1,170.3 | 415.0 | 526.1 | 43.1 |
| 1970 | 4,956.5 | 2,655.5 | 1,284.6 | 1,557.9 | −541.5 |
| 1974 | 7,500.6 | 3,769.5 | 1,726.3 | 2,492.4 | −487.6 |
| | | Per Annum Growth Rates (%) | | | |
| 1960–70 | 8.3 | 8.2 | 11.3 | 10.8 | — |
| 1970–74 | 10.6 | 8.8 | 7.4 | 11.8 | — |
| 1960–74 | 8.9 | 8.4 | 10.2 | 11.1 | — |

*Source:* U.S. Bureau of the Census, *City Government Finances*, various issues.

more from the fact that it subsidizes a broad array of functions than from the fact that the employees are gaining very high wages for performing normal public services.

This point is reinforced by examining growth rates over the 1960–74 period, as is done in Table 3. Over this time the net marginal functions deficit has grown by 11.1 percent per annum, more than expenditures either on the normal functions or on schools. Normal expenditures have grown by 8.4 percent, roughly 2.1 percent of which is a growth in employment, 2.4 percent is a growth in real wages per employee, and 4.0 percent is a growth in prices in New York City. School expenditures have grown by 10.3 percent—3.7 percent employment, 2.6 percent real wages, and 4.0 percent prices.[6]

### III. Comparison with Other Cities

It is also instructive to compare the New York City budget with that of the twenty nine other largest cities. This comparison can illustrate first the degree to which New York City is unique, and secondly the risk that the New York fiscal disease may spread to other cities.

Budget data for the general governments of the thirty largest *U.S.* cities for 1973–74 are given in Table 4. The budgetary items are grouped into normal and marginal functions categories, as was done in the bottom panel of Table 2, and are all expressed in per capita terms. Per capita expenditures on schools in the city are also shown in the table, even though New York City, Washington, Buffalo, Baltimore, Boston, and Memphis are the only other cities where school expenditures are actually in the general government budget.[7]

The table indicates first that only two large cities had current account deficits in 1973–74—New York City and Washington. But where New York City had to finance its deficit by borrowing from the bond market, Washington did not at the time have home rule and could borrow with federal guarantees (though this situation has changed as of January 1, 1976). Of the other cities, only Philadelphia had

[6] If the city's contribution to the pension fund were added into wages, the growth in real wages per employee would be slightly higher—2.6 percent and 2.8 percent respectively for normal and school categories.

[7] In the remaining cities schools are financed by an independent school authority. The data for schools comes from a different Census source that is not yet published for fiscal 1973–74. Thus data for the previous fiscal year are used throughout. Note that Honolulu has no local expenditures of any sort for schools because in Hawaii the schools are run by the state government.

TABLE 4—COMPARATIVE BUDGETS, THIRTY LARGEST U.S. CITIES, FISCAL 1973–74
(Dollars per capita)

| | Normal Revenue | Normal Expen-ditures | School Expen-ditures | Marginal Functions Deficit | Current Account Surplus |
|---|---|---|---|---|---|
| New York City | 962.1 | 493.0 | 207.0 | 325.9 | −63.8 |
| Buffalo | 660.3 | 318.0 | 272.0 | 17.1 | 53.2 |
| Baltimore | 796.1 | 358.6 | 207.9 | 40.6 | 189.0 |
| Boston | 835.8 | 487.3 | 188.1 | 117.4 | 43.0 |
| Philadelphia | 637.4 | 369.5 | 202.5 | 54.0 | 11.4 |
| Pittsburgh | 523.9 | 233.9 | 220.6 | 13.7 | 55.7 |
| Washington | 1,223.4 | 823.1 | 227.7 | 324.1 | −151.5 |
| Chicago | 512.2 | 247.1 | 219.0 | 20.9 | 25.2 |
| Cleveland | 628.7 | 350.7 | 214.0 | 8.3 | 55.7 |
| Columbus | 423.8 | 225.6 | 160.2 | 0.5 | 37.5 |
| Detroit | 702.9 | 334.2 | 219.9 | 98.4 | 50.4 |
| Indianapolis | 450.4 | 177.7 | 191.4 | 40.4 | 40.9 |
| Milwaukee | 564.2 | 260.9 | 217.0 | 25.3 | 61.0 |
| St. Louis | 572.3 | 299.1 | 173.3 | 60.3 | 39.6 |
| Kansas City | 568.3 | 292.1 | 173.4 | 38.0 | 64.8 |
| Atlanta | 674.8 | 294.8 | 204.2 | 25.9 | 149.9 |
| Jacksonville | 561.6 | 208.4 | 199.8 | 27.7 | 125.7 |
| Memphis | 489.8 | 204.4 | 168.0 | 20.4 | 97.0 |
| New Orleans | 487.7 | 253.0 | 138.8 | 24.1 | 71.8 |
| Dallas | 474.5 | 236.4 | 167.9 | 13.1 | 57.1 |
| Houston | 366.9 | 178.1 | 158.1 | 8.7 | 22.0 |
| San Antonio | 342.0 | 128.8 | 136.5 | 6.1 | 70.6 |
| Denver | 723.0 | 350.1 | 208.0 | 50.3 | 114.6 |
| Los Angeles | 607.4 | 242.5 | 224.7 | 50.9 | 89.3 |
| San Francisco | 956.4 | 466.2 | 194.7 | 225.0 | 70.5 |
| San Diego | 467.3 | 181.2 | 217.9 | 23.4 | 44.8 |
| San Jose | 492.4 | 163.1 | 282.0 | 14.4 | 32.9 |
| Phoenix | 480.1 | 210.0 | 208.8 | 5.9 | 55.4 |
| Seattle | 672.7 | 314.9 | 250.9 | 36.7 | 70.2 |
| Honolulu | 304.6 | 242.5 | — | 2.5 | 59.6 |

*Source:* U.S. Bureau of the Census, *City Government Finances*, 1973–74.

TABLE 5—NEW YORK CITY BUDGET IN RELATION TO THAT OF 27 OTHER LARGE CITIES
(Dollars per capita)

| | Normal Revenue | Normal Expen-ditures | School Expen-ditures | Marginal Functions Deficit | Current Account Surplus |
|---|---|---|---|---|---|
| New York City | 962.1 | 493.0 | 207.0 | 325.9 | − 63.8 |
| Mean of 27 cities[a] | 580.5 | 273.6 | 200.8 | 39.5 | 66.6 |
| *NYC*—Mean | 381.6 | 219.4 | 6.2 | 286.4 | −130.4 |
| New York City | 962.1 | 493.0 | 207.0 | 325.9 | − 63.8 |
| Unconstrained regression prediction[a] | 681.1 | 372.0 | 213.3 | 97.8 | − 2.0 |
| *NYC*—Reg. Pred. | 281.0 | 121.0 | −6.3 | 228.1 | − 61.8 |
| New York City | 962.1 | 493.0 | 207.0 | 325.9 | − 63.8 |
| Constrained regression prediction[a] | 758.3 | 361.5 | 225.9 | 98.6 | 72.3 |
| *NYC*—Reg. Pred. | 203.8 | 131.5 | −18.9 | 227.3 | −136.1 |

*Source:* U.S. Bureau of the Census, *City Government Finances*, 1973–74.

[a] Excluding New York City, Washington, and Honolulu from sample.

a current surplus that was small enough relative to revenues that it would have been converted to a deficit had the city's 1973–74 unemployment rate been as high as it is now.

In terms of composition, it is again the marginal functions where New York City stands out. The net deficit on these marginal functions was $326 per capita in New York City, roughly the same as Washington, which is also a state government and therefore has a large budget for welfare, higher education, and public hospitals. The only other cities that have marginal functions deficits of more than $60 per capita are San Francisco (large transit subsidies and pension contributions), Boston (large pensions and public housing), and Detroit (large pensions).

Table 5 makes this comparison more systematically. The top panel compares New York's per capita budget figures with the mean for the twenty-seven other large cities (all those listed in Table 4 except New York, Washington, and Honolulu). New York's normal expenditures are $219 more than average, school expenditures are above average, and the marginal functions deficit is $286 more than average. Revenues are fortunately also $382 more than average, but the current account deficit is still $130 more than average.

It can be argued that these comparisons are misleading because New York City is not an average large city. Its racial composition, income levels, employment growth, and age are not extraordinary as cities go, but its population size and density clearly are (the population density in New York City of 25,000 per square mile is 2.5 times as great as the next highest). One way to correct for those factors in making budgetary comparisons is to run regressions for the twenty-seven city sample, explaining budgetary variables as a function of independent variables such as population size and density, income and poverty levels, federal and state grants, and regional and governmental structure dummy variables. Values for New York City can then be inserted in the regressions and budgetary totals "predicted" for a hypothetical city with all of New York's properties.

When this comparison is made, it cuts down but does not eliminate the deviations of New York's actual budgetary values from their predicted values. Information in the second panel of Table 5, based on regressions explaining the four right-hand budgetary variables, leads to a positive residual of $121 for normal expenditures, $228 for marginal functions, and $62 for the current account deficit. Information based on regressions for all five budgetary variables, estimated in a way that insures the budget identity is preserved, leads to positive residuals of $132 for normal expenditures, $227 for marginal functions, and $136 for the current account deficit.[8] However this comparison is made, New York runs a substantially greater current account deficit than is predicted on the basis of other cities, has somewhat greater expenditures on normal functions, virtually no greater expenditures on schools, and much greater deficits for marginal functions.

## IV. The Bond Ratings

One of the interesting sidelights in the gradual progression of the city toward default was that an investor looking solely at the ratings of New York City bonds would have gotten rather misleading information. Back in the early 1960's, when the city first experimented with current account deficit spending, the city's bonds were rated *A* by both Moody's and Standard and Poor's Investors Services—of "upper medium" quality. In 1965–66, when the deficits had persisted for five

[8] The two sets of regressions are reported in Appendix A.

years, both services downgraded the city's bond rating to *Baa*—of "lower medium" quality (the classification is labelled *BBB* by Standard and Poor's, but it means the same as *Baa* for Moody's). This caused a 50 to 75 basis point increase in borrowing costs and raised a continuing popular outcry by city officials and congressmen against the private rating services, even despite the fact that the market itself was at this time forcing New York to borrow at much higher rates than other *Baa* borrowers.[9] The lengthy political campaign eventually seemed to pay off, however, and in December 1972, Moody's upgraded the city bonds again to *A*, despite the fact that the very large current account deficits were just then beginning. Standard and Poor's behaved even more curiously, holding out against upgrading the city until December 1974, but then upgraded the city's bonds to *A* just as the early signs of a potential default were being raised.[10]

The rating services are private agencies, and though they publish voluminous information on the budget of borrowing governments, it is impossible to tell exactly how they arrive at their ratings. A statistical attempt to explain their ratings by Carleton and Lerner turned up population size, the debt-assessed value ratio, the average tax collection rate, and whether or not the borrowing agency was a school district as statistically significant independent variables, but still left a large random component to the ratings.[11] The school dummy and population variables were only included to measure the lower rating given smaller and more unknown borrowers—if the examination was confined to a sample of large city general governments, the ratio of gross outstanding debt to the assessed value of property in the district would be the only controlling factor. While it is certainly one good indication of credit worthiness, this measure is not fully adequate for several reasons—it contains no adjustment for grants from state or the national government, no adjustment for changes, income, or sales taxes, no account of the ability of the city to alter tax rates on assessed value, and it only recognizes indirectly whether the city is or is not actually running current account deficits.

It is possible to estimate whether, or when, Moody's was being fair to New York City by their own standards in much the same way as before. A regression was fit to data from the 19 other largest cities (excluding New York and Washington), using pooled cross-section data for 1955, 1965, 1970, and 1974. The dependent variable was a rating index (1 for *Baa,* 2 for *A*, 3 for *Aa*, and 4 for *Aaa*) and the independent variable was the debt-assessed value ratio. The results, in Table 6, indicate that *by their standards* Moody's did indeed appear to be penalizing the city in the 1965–72 period—the "predicted" rating throughout this time was *A*. In retrospect, of course, this penalty was quite deserved. However, even *by their standards,* the value of New York City securities began dropping in 1972, just when the city's rating was upgraded, and this *A* rating was retained right up until the city could no longer market any securities in April 1975. This shift in the rating was difficult to justify then, impossible now.

The more fundamental question, of course, regards the standards. It is difficult and possibly unfair to make a very

[9] In 1971 New York was forced to pay 7.75 percent for twenty-year *Baa* bonds, while Greenwood, Mississippi, paid only 5.8 percent for twenty-year bonds of the same quality. A fascinating account of this whole experience is contained in a "Background Paper" by John E. Petersen.

[10] This situation is described in more colorful terms by a news release of Senator Thomas Eagleton, November 26, 1975.

[11] See William T. Carleton and Eugene M. Lerner. Their results are corroborated in broad detail by other studies (see Petersen, Ch. 7).

TABLE 6—PREDICTED AND ACTUAL RATINGS, NEW YORK CITY BONDS
(Current dollar totals in millions)

| Date | Gross Debt Outstanding | Assessed Value of Property | Ratio of Debt to Assessed Value | Actual Rating[a] | Predicted Rating[b] | Predicted Less Actual |
|---|---|---|---|---|---|---|
| 1955 | 4,845.8 | 20,277.8 | .239 | 2.0 | 2.30 | 0.30 |
| 1965 | 7,459.2 | 29,752.7 | .251 | 1.0 | 2.27 | 1.27 |
| 1970 | 8,690.8 | 34,292.3 | .253 | 1.0 | 2.27 | 1.27 |
| 1974 | 13,508.7 | 38,529.2 | .351 | 2.0 | 2.05 | 0.05 |

*Source:* U.S. Bureau of the Census, *City Government Finances;* various issues. Moody's Rating Service, various issues.

[a] Scale used is as follows:
*Baa* = 1.0
*A* = 2.0
*Aa* = 3.0
*Aaa* = 4.0

[b] On basis of the regression:

$$\text{Predicted Rating} = 2.824 - \underset{(1.6)}{2.197}\,(\text{Debt/Ass. Val.}), \qquad R^2 = .03$$
$$SE = .85.$$
sample size = 76

The dependent variable was measured in the scale used in footnote a, and the regression was fit to pooled cross-section observations on the nineteen largest cities, excluding New York City and Washington, in 1955, 1965, 1970, and 1974.

forceful criticism of the rating services—no general obligation bonds rated by the services have defaulted in the postwar period, and their information and stamp of approval has surely played a role in that good record. As the New York City case illustrates, however, there is always a first time, and in retrospect the information provided by the rating services in this particular case was not so good. They correctly penalized the city in the 1965–72 period, but incorrectly backed down in 1972. This faulty information may reflect a possible weakness in the criteria apparently used by the rating services—specifically, the fact that the deficit position of the current account budget appears to have little or no direct role in the ratings—or it could also reflect the possibility that the ratings may have been influenced by political pressures from city officials.

## V. Impact on the Economy

The present fiscal difficulties of New York and other cities arise in part because of the economic recession, but there is also some reverse feedback. If the city defaults, or flirts with default, this may set in motion a chain of events that will threaten the recovery from the recession. It then becomes important to try to determine whether there are any such multiplier effects for the New York situation.

The most immediate problem is the capital losses on bonds in default. If the city were to be unable to meet all its obligations, in all likelihood bondholders would be the ones to take the first losses. A large share of the outstanding bonds are held by banks in New York, and there was for a time thought to be a risk that the losses would be so great as to trigger a flight of deposits from the city banks. Most of the risk from this quarter seems to have passed, however. It never was highly likely,[12] and now that the federal government has agreed to make short-term loans to the city, and the Federal Reserve to make discount loans to any banks in difficulty, the risk of a chain reaction col-

[12] See the Congressional Budget Office Report, p. 19, and Robert Samuelson, "Troubled Friend at Chase Manhattan," *The New Republic,* November 15, 1975.

lapse of banks because of the New York fiscal crisis is relatively slight.

A second possible way in which the New York fiscal situation could retard the recovery is through the direct cutbacks or tax increases required to balance the city's budget. The New York State Emergency Financial Control Board plan for balancing the city's budget calls for expenditure cutbacks of about $100 million in fiscal 1975–76, $500 million in fiscal 1976–77, and $700 million in fiscal 1977–78. This comes on top of a previous 6 percent expenditure cutback and 7 percent revenue increase already made by the city, and a further $200 million tax increase as part of the package which induced the Administration to agree to short term federal loans to the city. New York State, with its own financial problems, has also recently raised taxes by $600 million. The net effect of all of these actions seems likely to reduce the aggregate demand stimulus contributed by state and local governments by about $2 billion, leading to a reduction in *GNP* of approximately $4 billion below what would otherwise be the case.[13]

A third factor that must be considered is the impact on the state and local bond rate. This impact can be tested as follows. A regression was fit of the form

$$CORP\ (1-t) - RATE = a_0 + a_1 U + a_2\, TIME,$$

where *CORP* is the corporate *Aaa* rate, *RATE* is the municipal bond rate in the appropriate risk class (*Aaa, Aa, A,* or *Baa*), $t$ is the estimated average personal and corporate marginal tax rate among state and local bondholders, $U$ is the unemployment rate, and *TIME* is a time trend. The regression explains the rate differential between the after-tax corporate bond rate and the relevant municipal rate as a function of the unemployment rate (since state and local governments must borrow relatively more in a recession, their rate would rise relative to the corporate rate) and time (since municipal rates have been falling relative to the corporate rate over time). The equation was fit to halfyear data from 1955-74, and then extrapolated to 1975 to see how values predicted by the regression compare with actual municipal rates in 1975.

The comparisons are given in Table 7. For the *Aaa* rate, the actual municipal rate has averaged 13 basis points above the predicted rate throughout 1975, 20 basis points in the third quarter. For bonds in the more risky categories, the residuals are much larger, rising to a total of 103 basis points in the third quarter for *A* bonds and 96 for *Baa* bonds. This difference in risk spread demonstrates what Treasury Secretary William E. Simon calls a "flight to quality" in these hard times.[14] The problem, of course, is that in many cases the lack of quality might not be a function of the locality's own fiscal behavior but reflect the impact of outside uncontrollable forces such as the economic recession.

It is not clear whether the entire residual in 1975 actual municipal bond yields can be blamed on New York City, but it seems likely that a large component can. The residual is greatest in the higher risk borrowing classes the city is typically classified in and has become greatest when the news about the city's fiscal problems was the most alarming. If we make the extreme assumptions that the entire residual is caused by the New York situation and that the corporate rate is not *lowered* by

[13] These estimates accord with those of the Joint Economic Committee, p. 55.

[14] See Simon's statement of September 24, 1975. Among other things, the statement is ironic for the degree to which it downplays the risk that state and local interest rates will rise for high quality borrowers. Simon spent the first half of the year warning of the danger that federal deficits will raise interest rates for all borrowers and crowd them out of credit markets.

TABLE 7—PREDICTED AND ACTUAL STATE AND LOCAL BOND RATES, 1975
(In percentage points, by months)

| Month | *Aaa* Bonds Actual | *Aaa* Bonds Predicted[a] | *Aaa* Bonds Difference | *Aa* Bonds Actual | *Aa* Bonds Predicted[a] | *Aa* Bonds Difference | *A* Bonds Actual | *A* Bonds Predicted[a] | *A* Bonds Difference | *Baa* Bonds Actual | *Baa* Bonds Predicted[a] | *Baa* Bonds Difference |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Jan. | 6.39 | 6.25 | 0.14 | 6.57 | 6.42 | 0.15 | 7.13 | 6.55 | 0.58 | 7.45 | 6.76 | 0.69 |
| Feb. | 5.96 | 6.08 | −0.12 | 6.11 | 6.24 | −0.13 | 6.53 | 6.38 | 0.15 | 7.03 | 6.59 | 0.44 |
| March | 6.28 | 6.12 | 0.16 | 6.49 | 6.29 | 0.20 | 6.79 | 6.42 | 0.37 | 7.25 | 6.63 | 0.62 |
| April | 6.46 | 6.34 | 0.12 | 6.82 | 6.51 | 0.31 | 7.09 | 6.64 | 0.45 | 7.43 | 6.85 | 0.58 |
| May | 6.42 | 6.29 | 0.13 | 6.76 | 6.47 | 0.29 | 7.13 | 6.59 | 0.54 | 7.48 | 6.81 | 0.67 |
| June | 6.28 | 6.19 | 0.09 | 6.72 | 6.35 | 0.37 | 7.36 | 6.48 | 0.88 | 7.49 | 6.70 | 0.79 |
| July | 6.39 | 6.24 | 0.15 | 6.80 | 6.40 | 0.40 | 7.50 | 6.54 | 0.96 | 7.60 | 6.74 | 0.86 |
| Aug. | 6.40 | 6.33 | 0.07 | 6.80 | 6.49 | 0.31 | 7.55 | 6.62 | 0.93 | 7.71 | 6.83 | 0.88 |
| Sept. | 6.71 | 6.33 | 0.38 | 7.10 | 6.49 | 0.61 | 7.81 | 6.62 | 1.19 | 7.96 | 6.84 | 1.12 |
| QI | 6.21 | 6.15 | 0.06 | 6.39 | 6.32 | 0.07 | 6.82 | 6.45 | 0.37 | 7.24 | 6.66 | 0.58 |
| QII | 6.39 | 6.27 | 0.12 | 6.77 | 6.44 | 0.33 | 7.19 | 6.57 | 0.62 | 7.47 | 6.79 | 0.68 |
| QIII | 6.50 | 6.30 | 0.20 | 6.90 | 6.46 | 0.44 | 7.62 | 6.59 | 1.03 | 7.76 | 6.80 | 0.96 |
| 9 mo. avg. | 6.37 | 6.24 | 0.13 | 6.68 | 6.41 | 0.27 | 7.21 | 6.54 | 0.67 | 7.49 | 6.75 | 0.74 |

[a] From the equation

$$CORP\ (1 - t) - RATE = a_0 + a_1 U + a_2\ TIME,$$ where *CORP* is the *Aaa* corporate rate, *RATE* stands for the four risk classifications above, $U$ is the unemployment rate, *TIME* is a time trend, and $t$ is an estimated average income tax rate. The equation was fit by half years from 1955–74. In all categories a value of $t=.2$ gave the lowest standard errors. The estimated coefficients (with $t$-ratios in parentheses) were as follows:

| RATE | $a_0$ | $a_1$ | $a_2$ | $R^2$ | $SE$ |
|---|---|---|---|---|---|
| *Aaa* | .021 | .003 (.1) | .038 (7.2) | .58 | .19 |
| *Aa* | −.096 | −.014 (.4) | .042 (7.5) | .78 | .21 |
| *A* | −.424 | — | .046 (7.6) | .78 | .23 |
| *Baa* | −.857 | −.015 (.4) | .064 (9.1) | .83 | .26 |

these troubles in the municipal market, the New York situation will reduce the extent to which state and local governments borrow to finance capital construction projects. Weighting the third quarter residuals by the proportion of borrowing in each risk class leads to an increase in the state and local bond rate of 63 basis points, which should by itself reduce state and local construction by about $4.5 billion.[15] The *GNP* impact of this would be about $9 billion in the first year or two, and when added to the impact of the direct cuts discussed above, both together would reduce *GNP* by .7 percent and raise the unemployment rate by about .25 percentage points over what it would otherwise be.[16] This is quite a large effect for the fiscal difficulties of just one city, large enough by itself to offset the tax cut recently passed by Congress.

Moreover, focusing only on this surprisingly large macroeconomic impact can obscure two points. The first is that if the city does default, and if that default does threaten the solvency of major New York

[15] Using an estimate derived in Gramlich and Galper, Table 4. This estimated construction effect is about the same as that given by F. Gerard Adams and James N. Savitt, though they anticipate a larger rise in state and local interest rates and use a smaller interest rate elasticity for construction.

[16] The interest rate effects are larger than those found by Ronald W. Forbes and John E. Petersen, because I have allowed the municipal *Aaa* rate to be affected by the New York City problem. I have not included observations from 1975 in the regression (as they did), and I have worked in the impact of the tax exemption of state and local interest payments (as they did not). The analyses also differ in that my calculation focuses on the probable effect on state and local borrowing and investment, whereas theirs focuses on the rise in interest payments, assuming that borrowing is unaffected.

banks and cause a loss of deposits, the situation becomes much more uncertain and the risk of a general collapse of borrowing and investment much greater. In this context, one great advantage of bailing out the city is precisely this—it keeps the general economic risks predictable and manageable.

The second point is that this estimate of the impact of any New York default is only manageable for the country as a whole—the impact on the economy and the citizens of the city itself will be extremely painful. Table 5 showed that per capita taxes are already much higher in New York than in other cities, and the measures taken to put the city on a sounder fiscal footing will raise them further, increasing the risk of a flight of the taxpaying businesses and households to other localities. On the expenditure side, the *MAC* figures project that by fiscal 1978 55 percent of the city's expenditures will be "uncontrollable"—mandated for interest, debt service, welfare, and pension payments. This forces very deep cutbacks in the remaining controllable portion, estimated by the Joint Economic Committee to be 18 percent in real terms. Even with a freeze on money wages, these are severe cutbacks. And even these severe cutbacks might not be adequate to balance the city's budget by 1978 if they lead to other reductions in federal and state grants mandated for the programs cut back or if they reduce levels of economic activity and revenues in the city.

## VI. What is to be Done in the Short Run?

It is generally easier to analyze what went wrong than to say what to do about it. Several principles do emerge from the preceding facts, however, and I will try to highlight the pertinent ones.

The first is that the city itself is ultimately responsible for its financial woes—the budget deficits started long ago, in part because of official underestimates of the severity of the problem; the deficits were large even before the current recession, and not caused by any one particular program but a host of them. The second is that even though the problem is of the city's own making, things have reached such a point that now the city simply cannot correct the problem by itself—tax burdens in the city and state are already extremely high, uncontrollable expenditures are extremely high, and the cash expenses necessary to roll over the short-term debt when the market will not are excruciating. Even with the severe cuts in expenditures mandated by the Emergency Financial Control Board, it will take the city three years to restore balance in its budget, and only then under relatively optimistic assumptions regarding future revenues.

The question facing the national government is whether to take steps to avoid the city's having to default on the bonds it must retire in the near future.[17] If this were not done, there is at least some chance that a financial panic could develop, and some chance that the economic recovery could be aborted even if a panic did not develop. There is also an excellent chance that the delivery of vital services in the city—police, fire, sanitation, health, education—could be temporarily or permanently disrupted by default, and that it would be very difficult to pick up the pieces. Minimizing each of these fears is essentially the rationale for not allowing the city to default.

On the other side, the advantage of a default is that the city has in fact signed certain contracts that it is simply unable to meet, it has to improve its budget pro-

[17] Default in this sense means having the court, and not the Emergency Financial Control Board, take over the reconciliation of inconsistent claims on the city. In a narrow legal sense, one could argue that the recent state action to substitute *MAC* bonds for some of the maturing city debt implies that the city is already in default.

cess and reduce its deficits, and it never will unless it is forced to go through the default process. The last assumption is the key one—by this time the state itself and through its Emergency Financial Central Board has exerted so much control on the city that it effectively lost its budgetary autonomy and will have to abrogate some contracts, whether or not there is a default. Given this, these marginal benefits of a default are probably unimportant and it seems almost foolish to risk the possible severe dangers of allowing the city to go under. By this time, all but the most hard-nosed conservatives appear to agree with this position.

## VII. Long-Term Measures

The city's plight does illustrate several more long-term issues, however, and it is well to emphasize them before such crises recur. The first is that better state and local financial accounting is urgently needed. There is no obvious reason why the budgets of large state and local governments could not be more quickly made available to the financial community, tabulated in standard form, using official conventions on what is current and what capital spending, on the increase in pension fund liabilities, on the size and maturity of outstanding debt. Whether the bond rating authority is then brought under the Securities and Exchange Commission, as Senator Eagleton wishes, there should also be much more explicit statement of why bonds are rated as they are, the criteria used, and the data behind these criteria.[18] As a specific suggestion, the rating agencies have apparently relied solely on stock figures such as the ratio of debt to assessed property values; it would seem that flow figures such as the recent size of current account surplus, the controllable portion of the budget, and the dangers in not being able to roll over short term debt should also enter in.

A second point regards the intrinsic financial difficulties of state and local governments when macroeconomic conditions change. State and local revenues are now composed largely of income and sales taxes—those two sources comprised 60 percent of own revenue in 1973—and net expenditures for welfare and medicaid have also become very large. When aggregate conditions change, state and local budgets will tend to operate in an automatically stabilizing manner just as will the federal budget. The only difference is that the inability to run current account deficits at the state and local level forces the stabilizers to be offset by other spending costs or tax increases. It can, of course, be argued that governments should have had the foresight to build up a stock of liquid assets for just such an emergency. This argument is not terribly persuasive in a time like the present, however, when the business cycle decline is so much steeper than any government could possibly have anticipated. In any case to illustrate the problem, a special Joint Economic Committee survey of eighteen large cities found only five that did not raise taxes or cut spending in the past year.[19] This is both counterproductive from a macro standpoint and, more importantly, it leads to instability in the delivery of services at the state and local level. A sensible remedy would be to give out revenue sharing money on a countercyclical basis, as Senator Muskie wants to do. It was emphasized above that the recession was not at the heart of New York's woes, and these woes would have only been modestly ameliorated by a countercyclical revenue sharing program, but they would have been ameliorated and the program would still seem to have appeal in a cyclical economy.

[18] See, for example, the Twentieth Century Fund.

[19] Joint Economic Committee, p. 29.

A final point involves the redistributive nature of the city's social programs. As was argued earlier, New York's difficulties can in large part be traced to the fact that it was doing too much on its own. In some cases—pension contributions, higher education, transit subsidies, public hospitals—the choice was the city's own; in others—welfare and medicaid—the choice was made by the state and mandated on the city's budget. Laudable as these choices may be from a social standpoint, local governments do have to be careful in a federal system, first, not to laden taxpayers with too much debt and forced debt service payments, and, second, not to undertake too much redistribution at the local level. If these programs are to be done, they must at least be financed on a national scale. In this sense, the New York experience can also be interpreted as something of a setback for the new federalism, which some advocates interpret as turning functions such as income redistribution back to the states and localities.

TABLE A.1—UNCONSTRAINED REGRESSIONS PREDICTING BUDGETARY VARIABLES
($t$-ratio below coefficient)[a]

| Dependent Variable | Intercept | Population | Disposable Income | Poverty | Nonwhite | Normal Grants | Density | Midwest | South | City-County | $R^2/SE$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Normal expenditures | −426.0 | −29.6 | .075 | 1,163.9 | — | .472 | 12.3 | — | — | — | .665 |
| | | (−1.5) | (3.1) | (2.7) | | (1.8) | (3.3) | | | | 56.5 |
| School expenditures[b] | 317.0 | — | — | −525.5 | 56.1 | — | — | −42.4 | −44.0 | — | .689 |
| | | | | (−4.0) | (1.3) | | | (−3.9) | (−4.0) | | 21.3 |
| Marginal functions deficit | −356.8 | −13.0 | .059 | 468.5 | — | — | 5.4 | — | — | 27.2 | .657 |
| | | (−1.3) | (4.6) | (2.1) | | | (3.1) | | | (2.0) | 29.9 |
| Current account surplus | 42.7 | — | — | — | — | .576 | −6.0 | — | — | — | .619 |
| | | | | | | (5.8) | (−4.4) | | | | 25.6 |
| Normal Revenue[c] | −423.1 | −42.6 | .134 | 1,106.9 | 56.1 | 1.048 | 11.7 | −42.4 | −44.0 | 27.2 | — |

*Source:* U.S. Bureau of the Census, *City Government Finances;* Sales Management, *1974 Survey of Buying Power;* U.S. Dept. of Commerce, *Statistical Abstract of the U.S., 1973.*

[a] Each dependent variable is in per capita terms. Population is the gross size of the city's 1973 population in billions, Disposable Income (*DI*) is the average disposable income for the city in per capita terms in 1973, Poverty is the 1973 ratio of households with *DI* below $5000 to the total number of households in the city, Nonwhite is the 1973 ratio of the nonwhite population to the total, Normal Grants is the per capita total for normal functions and untied in 1973, Density is the number of thousands of people per square mile in the city, all the remaining variables are one for the indicated category and zero otherwise.

[b] Using slightly different data, where the first five variables are defined for the county area in accordance with school district boundaries.

[c] Implied by the summation of coefficients in the other four rows.

TABLE A.2—CONSTRAINED REGRESSIONS PREDICTING BUDGETARY VARIABLES
($t$-ratio below coefficient)[a]

| Dependent Variable | Intercept | Population | Disposable Income | Poverty | Normal Grants | Density | Midwest | South | City-County | $R^2/SE$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Normal expenditures | −429.8 | −25.6 | .074 | 1,236.7 | .445 | 10.0 | 16.4 | −21.0 | 15.2 | .682 |
| | | (−1.2) | (2.7) | (2.5) | (1.6) | (2.0) | (0.6) | (−0.5) | (0.5) | 59.5 |
| School expenditures | 291.3 | −2.0 | −.004 | −336.8 | .223 | 1.2 | −31.7 | −32.9 | −28.9 | .666 |
| | | (−0.2) | (−0.4) | (−1.7) | (2.0) | (0.6) | (−2.6) | (−1.9) | (−2.5) | 24.4 |
| Marginal functions deficit | −380.7 | −12.3 | .062 | 436.2 | .117 | 5.1 | 11.3 | 9.3 | 25.6 | .677 |
| | | (−1.1) | (4.3) | (1.7) | (0.8) | (1.9) | (0.7) | (0.4) | (1.7) | 31.3 |
| Current account surplus | 41.4 | 11.1 | −.003 | 44.0 | .557 | −6.5 | −11.8 | 1.2 | 10.4 | .684 |
| | | (1.1) | (−0.3) | (0.2) | (4.5) | (−2.9) | (−0.9) | (0.1) | (0.8) | 26.9 |
| Normal revenue | −477.8 | −28.8 | .129 | 1,380.1 | 1.342 | 9.8 | −15.8 | −43.4 | 22.3 | .797 |
| | | (−1.0) | (3.6) | (2.2) | (3.8) | (1.5) | (−0.4) | (−0.8) | (0.6) | 77.3 |

*Source:* U.S. Bureau of the Census, *City Government Finances;* Sales Management, *1974 Survey of Buying Power;* U.S. Dept. of Commerce, *Statistical Abstract of the U.S., 1973.*

Specification is the same as in Table 6, except that Nonwhite variable is dropped from all equations. Every other independent variable is entered in all equations to insure consistency.

## REFERENCES


F. G. Adams and J. N. Savitt, "Statement of the House Budget Committee," Oct. 23, 1975.

W. T. Carleton and E. M. Lerner, "Statistical Credit Scoring of Municipal Bonds," *J. Money, Credit, Banking,* Nov. 1969.

Congressional Budget Office, "New York City's Fiscal Problem, Its Origin, Potential Repercussions, and Some Alternative Policy Responses," Oct. 10, 1975, 13, 19.

R. W. Forbes and J. E. Petersen, "Costs of Credit in the Municipal Bond Market," Municipal Finance Officers Association 1975.

E. Gramlich and H. Galper, "State and Local Fiscal Behavior and Federal Grant Policy," *Brookings Papers,* 1973, *1.*

R. D. Horton, *Municipal Labor Relations in New York City: Lessons of the Lindsay-Wagner Years,* 1973.

J. E. Petersen, *The Rating Game,* Twentieth Century Fund 1974, 125.

Sales Management, *1974 Survey of Buying Power.*

R. Samuelson, "Troubled Friend at Chase Manhattan," *The New Republic,* Nov. 15, 1975.

U.S. Bureau of the Census, *City Government Finances,* 1973–74.

U.S. Dept. of Commerce, *Statistical Abstract of the U.S.,* 1973.

# The Effects of Family Background on Earnings

By Mary Corcoran, Christopher Jencks and Michael Olneck*

Men with socially and economically privileged parents usually earn more as adults than men with less privileged parents. There is no reason to suppose that men with privileged parents have a stronger preference for cash as against psychic income from their work. If anything, the contrary seems likely. It follows that men with privileged parents are either more valuable to "rational" employers, search more effectively for lucrative jobs, or benefit from positive discrimination by employers based on either background per se or its correlates. This paper tries to assess the importance of each of these factors.

Both the size and apparent causes of earnings differentials between men from different backgrounds depend to a considerable extent on how you define background. We will begin with a set of parental characteristics that are widely believed to enhance a child's life chances: being white, being born in the United States, having a lot of education, engaging in high-status occupations, having a substantial income, living in the North rather than the South, living in an urban area rather than on a farm, having relatively few children, and not being separated or divorced. The top rows of Table 1 list four national surveys of men aged 25–64 that provide data on most of these parental characteristics. The data cover only males.[1]

The standard deviation of the natural logarithm of earnings (LnEarnings) is higher in the first and last of these four surveys than in the two intervening surveys. This is due to methodological differences. The two intervening surveys were conducted by Survey Research Center (*SRC*) and cover only heads of households. This eliminates some very low earners. *SRC* also locates fewer household heads who report very low earnings. This may be because *SRC* at the University of Michigan has a lower response rate than the Census or Current Population Survey (*CPS*), or because Census and *CPS* respondents (who are often wives) understate the household head's earnings more than *SRC* respondents.

Almost all the parental advantages

* Assistant Professor of Political Science at Michigan State University, Professor of Sociology at Harvard University, and Assistant Professor of Education at the University of Wisconsin, respectively. This paper draws on data analyses conducted by Susan Bartlett, James Crouse, David Eaglesfield, Greg Jackson, Kent McClelland, Sherry Morgan, and Peter Mueser as well as ourselves at the Center for the Study of Public Policy in Cambridge, and on unpublished work by John Brittain of the Brookings Institution and Robert Hauser of the University of Wisconsin. We have received financial support from the Carnegie Corporation of New York, the Ford Foundation, the National Institute of Education, the Employment and Training Administration of the Department of Labor, the Office of Income Security Policy Research and Evaluation of the Department of Health, Education and Welfare. None of these individuals or organizations is responsible for our opinions or errors.

[1] David Featherman and Hauser report that background has virtually no impact on the earnings of married women who work.

TABLE 1—RELATIONSHIPS BETWEEN LNEARNINGS AND FAMILY BACKGROUND FOR CIVILIAN NONINSTITUTIONAL NONSTUDENT MALES WITH POSITIVE EARNINGS AND COMPLETE DATA

| Survey Organization (Year of Earnings Data)[a] (1) | Survey Name (2) | Additional Restrictions on Sample (3) | Age in Earning Year (4) | Sample Size (5) | Background Measures[b] (6) | Test Score (7) | S.D. of Ln Earnings (8) | S.D. of Predicted LnEarnings[c] ($R^2$'s are in parentheses) (9) | S.D. of Predicted Family Means[d] (Sibling $r$) (10) |
|---|---|---|---|---|---|---|---|---|---|
| *CPS* (1961) | Occupational Changes in a Generation (*OCG*) | Includes students and men with earnings = 0 (if income > 0) | 24–63 | 11,505 | 1, 2, 3, 4, 5, 7, 8, 9, 10 | No | .820[e] | .334[e] (.166) | |
| *SRC* (1964) | Productive Americans (*PA*) | Heads of households | 24–63 | 1,188 | 1, 2, 3, 7, 8, 9 | No | .707 | .309 (.191) | |
| *SRC* (1971) | Panel Study of Income Dynamics (*PSID*) | Heads of households | 25–64 | 1,774 | 1, 2, 3, 4, 5, 7, 8, 9 | Yes | .753 | .264 (.123) | |
| *NORC* (1973) | *NORC* Brothers | Brothers | 24–63 | 300 | 1, 3, 4, 5, 8, 9, 10 | No | .870 | .220 (.064)[g] | .312 (.129) |
| *CPS* (1966) | Parnes' National Longitudinal (*NLS*) | None | 45–59 | 2,580 | 1, 3, 4, 5, 7, 8, 10 | No | .883 | .343 (.151) | |
| Talent (1971–2) | Talent Twins and Siblings | Brothers who reached 11th or 12th grade | 28(±1) | 198 | 1, 3, 4, 9, 10 | Yes | .406 | .069[f] (.029) | .185[f] (.207) |
| Olneck (1973) | Kalamazoo Brothers | Brothers who reached 6th grade in Kalamazoo, Michigan | 35–59 | 692 | (1), 2, 3, 4, 5, 6, (7), (8), 9, 10 | Yes | .446 | .126 (.080)[g] | .209 (.220) |
| Sewell et al. (1967) | Wisconsin | Nonfarm men who reached 12th grade in Wisconsin | 27(±1) | 2,069 | 3, 4, 6, (7), (8), 11 | Yes | .532 | .051 (.009) | |

[a] *CPS* = Current Population Survey, U.S. Bureau of the Census; Talent = Project Talent, American Institutes for Research, Palo Alto; *SRC* = Survey Research Center, University of Michigan; *NORC* = National Opinion Research Center, Chicago.
[b] 1 = white, 2 = father born in United States, 3 = father's education, 4 = father's occupation (Duncan scale), 5 = father white collar, 6 = mother's education, 7 = son's region of birth, 8 = son raised on farm, 9 = number of siblings, 10 = son not living with father when 15, 11 = parental income at 17. Variables in parentheses have no variance due to sample restrictions.
[c] Values predicted using all background variables in column 6, including significant nonlinearities and interactions.
[d] (column 8) $(r_i)^{1/2}$, where $r_i$ = intraclass $r$ in parentheses.
[e] Ln Income.
[f] Ln Hourly Wage.
[g] These regressions include age. Without age $R^2$'s are .055 for Kalamazoo and .051 for *NORC*.

listed above had positive effects on men's earnings. The only surprise is that men with foreign-born fathers earn about 10 percent more than men with American-born fathers in the three surveys that measure father's place of birth. This advantage disappears once we control region of residence and community size, but even then no disadvantage emerges. Almost all the background traits had essentially linear effects on LnEarnings. The exception was family size. Additional siblings always lower a man's expected earnings, but the negative impact diminishes as the absolute number of siblings increases. While a few other nonlinearities were significant in at least one survey, none had a consistent sign across all the samples and none was significant in more than one.

Since different surveys measured different background characteristics, the multivariate results are not strictly comparable. But no multiplicative interaction between different parental characteristics was consistently significant across surveys, or even had a consistent sign. Taken together, nonlinearities and interactions never accounted for more than .6 percent of the variance in LnEarnings in any survey, so we will ignore them hereafter.

If one compares the first two national surveys of men aged 25–64 to the last two, the variance explained by background characteristics declines from a weighted average of 17 percent in 1961–

64 to 11 percent in 1971–73. Given the many differences between samples and variables, this change proves nothing in itself. But *CPS* also conducted a meticulous replication of the 1962 Occupational Changes in a Generation (*OCG*) survey in 1973. Featherman and Hauser, using untransformed earnings, report that for married men aged 25–64, the coefficient of variation was the same in 1972 as in 1961, but the percentage of variance explained by father's occupation, farm upbringing, and number of siblings fell by 40 percent. This suggests that we should not write off the decline in $R^2$ in Table 1 as a methodological artifact.

Column 9 of Table 1 shows the standard deviation of predicted earnings for each sample. This indicates the degree of inequality we might expect if the parental advantages and disadvantages measured in these surveys were the only source of inequality in sons' earnings. The standard deviation of predicted incomes was .334 in 1961. Men whose parents ranked in the bottom quintile earned 60.4 percent of the national average. Those whose parents ranked in the top quintile typically earned 1.48 times the national average. Assuming a 40 percent decline in the variance traceable to parental advantages between the early 1960's and the early 1970's, the standard deviation of predicted incomes should have been .258 in 1972. This is quite close to the observed standard deviation of predicted earnings for 1971, which was .264 in the Panel Study in Income Dynamics (*PSID*). It is higher than the 1973 estimate from the National Opinion Research Center (*NORC*) brothers survey, which was only .220. But the *NORC* brothers survey included only 300 respondents, so the discrepancy could easily be due to sampling error.

None of the national surveys listed in Table 1 measured parental income. The one survey that does include this information covers Wisconsin high school graduates. In that sample average parental income when the son was 17 to 21 (taken from state tax records) had a significant impact on the son's earnings at age 27; a 1 percent increase in the parents' income was associated with an .08 percent increase in son's earnings at 27 years.[2]

More accurate measurement of father's education, mother's education, and especially the father's job characteristics might also increase the apparent explanatory power of parental status. But we doubt that correcting these problems would raise the variance explained by these traits by more than 50 percent. It would have less effect on the explanatory power of traits like race, region of birth, farm upbringing, and number of siblings which are already quite accurately measured and account for as much of the variance in earnings as parental education and occupation. We therefore suspect that even if we measured the parental traits listed in Table 1 altogether accurately, we would not be able to explain more than about a sixth of the variation in earnings today.

But why should backgrounds explain even a sixth of the variance in earnings? Our data suggest that employers hire and promote largely on the basis of the applicant's own characteristics. Controlling test scores and educational attainment makes the coefficients of father's education, father's occupation, number of siblings, and coming from a broken home extremely small and usually insignificant. Southern birth and farm upbringing exert some effect even among men with similar test scores and education, but these effects become negligible once we control the size of the community and region in which the respondent now lives. This suggests that

[2] These results are based on unpublished materials supplied by William Sewell and Hauser.

employers pay very little attention to whether a man is southern or farm born. These men are handicapped partly because they get less education and have lower test scores, and partly because they usually remain in communities where everyone is paid less than in the urban North.

The only background characteristic that has a substantial effect on earnings with *everything* controlled is race. Race is, of course, not just a characteristic of the respondent's parents but a visible characteristic of the respondent himself. The fact that employers paid blacks less than whites with the same test scores and the same amount of education strongly implies that in this instance employers engaged in "pure" discrimination.

So far we have concentrated on the effects of measured parental advantages. If these were the only parental characteristics that affected a child's earnings, we would expect children who had had the same measured advantages to end up as much alike as children who were actually raised in the same home. To determine whether this is true, we have looked at several surveys of brothers. Table 1 describes the samples. Even allowing for a secular decline in the impact of parental status, measured background characteristics explain less of the variance in earnings in our three samples of brothers than in our larger national samples. In the case of the *NORC* brothers this is probably just a matter of random sampling error. In the case of the Kalamazoo brothers, the target population may be atypical. (Note the severely restricted variance of LnEarnings.) In the case of the Talent brothers, the respondents are only 28, and no variable correlates as well with earnings at 28 as later. Whatever its causes, the low explanatory power of measured background characteristics implies that the correlation between brothers in these three samples may be lower than it would be in a representative national sample. Our suspicions on this score are strengthened by the fact that Brittain of the Brookings Institution obtained an intraclass correlation of .436 between the earnings of 151 men from 66 families from the Cleveland area. If we ignore the Talent sample on the grounds that it is too young to prove much, and pool the *NORC*, Kalamazoo, and Brittain results, we get an estimated intraclass correlation between brothers' earning of .23, with a standard error of about $\pm$ .04.

Unless brothers exert a direct effect on one another's earnings, the intraclass correlation between their earnings is equal to the percentage of variance we would be able to explain if we regressed earnings on all the genetic and environmental factors that brothers have in common. The intraclass correlations are at least twice as large as $R^2$ in each of our samples of brothers. This suggests that unmeasured background characteristics exert as much impact on earnings as the measured characteristics discussed earlier. Note, however, that most of the variance in eventual earnings is still within families.

If the unmeasured family characteristics that produce resemblance between brothers were similar in character to the measured characteristics that produce such resemblance, we would expect the unmeasured characteristics to exert most of their influence indirectly, via test scores and educational attainment. The data in Table 2 show that this is not the whole story. If the *NORC* brothers' earnings were alike solely because their educational attainments were alike, for example, the correlation between their earnings would be only (.356) (.528) (.356) = .067, not .129. If test scores and education together accounted for resemblance between brothers' earnings, the correlation between Kalamazoo brothers would be .113, not .220. This means that

TABLE 2—CORRELATIONS BETWEEN BROTHERS' CHARACTERISTICS

| | $X$ | $Q$ | $U$ | $Y$ | $X'$ | $Q'$ | $U'$ | $Y'$ |
|---|---|---|---|---|---|---|---|---|
| | 1.000 | | | | | | | |
| $X$ | 1.000 | | | | | | | |
| | 1.000 | | | | | | | |
| | — | 1.000 | | | | | | |
| $Q$ | .260 | 1.000 | | | | | | |
| | .387 | 1.000 | | | | | | |
| | .339 | — | 1.000 | | | | | |
| $U$ | .383 | .576 | 1.000 | | | | | |
| | .371 | .632 | 1.000 | | | | | |
| | .124 | — | .355 | 1.000 | | | | |
| $Y$ | .200 | .366 | .409 | 1.000 | | | | |
| | .096 | .356 | .366 | 1.000 | | | | |
| | — | — | — | — | 1.000 | | | |
| $X'$ | .775[a] | .253 | .389 | .197 | 1.000 | | | |
| | .797[a] | .419 | .409 | .094 | 1.000 | | | |
| | — | — | — | — | — | 1.000 | | |
| $Q'$ | .253[a] | .469 | .400 | .169 | .260 | 1.000 | | |
| | .419[a] | .580 | .451 | .216 | .387 | 1.000 | | |
| | — | — | .528 | .171 | .339 | — | 1.000 | |
| $U'$ | .389[a] | .400 | .549 | .269 | .383 | .576 | 1.000 | |
| | .409[a] | .451 | .546 | .210 | .371 | .632 | 1.000 | |
| | — | — | .171 | .129 | .124 | — | .355 | 1.000 |
| $Y'$ | .197[a] | .169 | .269 | .220 | .200 | .366 | .409 | 1.000 |
| | .094[a] | .216 | .210 | .207 | .096 | .356 | .356 | 1.000 |

*Note: NORC* (top row), Kalamazoo (middle row), Talent (bottom row).
$X$ = respondent's report of father's occupational status at 15 (Duncan score)
$Q$ = test score prior to school completion
$U$ = Highest grade completed
$Y$ = LnEarnings
Primes denote the second member of a given pair. All correlations were computed from files in which every pair appears twice, with order reversed. This makes product-moment correlations equal to intraclass correlations.

[a] In a few cases where reports of father's occupation were missing, the report of a brother was substituted. Therefore, the correlation between $X$ and $X'$ is slightly higher than the reliability coefficient.

unmeasured background characteristics must affect earnings in some way that is independent of test scores and educational attainment.

We can think of at least two ways in which unmeasured family characteristics could affect men's earnings independently of test scores and educational attainment. One possibility is that certain families socialize their children and develop noncognitive skills which are largely independent of test scores and educational attainment but still valuable to employers. If we accept the notion that "like produces like," then children who have these interpersonal skills are likely to come from homes where parents also possess these skills. If employers value these interpersonal skills, then parental traits which produce these skills should be correlated with parental status.

Another possibility is that some fami-

lies encourage their children to maximize earnings, while others encourage their children to maximize other forms of psychic income, even at the expense of earnings. If this were the case, we would expect the unmeasured family characteristics that maximize earnings net of test scores and educational attainment (i.e., parental values) to be quite different from those that maximize test scores and educational attainment. Indeed, the correlation might well be negative.

If one constructs a hypothetical variable to account for resemblance between brothers' test scores, another to account for resemblance between their educational attainments (over and above what is expected on the basis of test score similarity), and a third to account for resemblance between earnings (over and above what is expected on the basis of similar test scores and education), the last of these variables has a very weak, sometimes negative correlation with the other two, also with father's occupation. This suggests that our hypothetical variable does not primarily measure parental noncognitive skills valuable to employers. Rather our hypothetical variable may to a large extent be a proxy for familial values, and may be picking up a preference for monetary rather than nonmonetary rewards.

Our analyses appear to support three conclusions:

(1) Employers do not often discriminate directly against employees on the basis of parental characteristics other than race.

(2) Parental characteristics exert considerable influence on children's test scores and educational attainment, which in turn influence their earnings.

(3) Unmeasured parental characteristics also exert considerable impact on son's earnings, independently of test scores and educational attainment. The unmeasured characteristics that create these effects are almost uncorrelated with measured advantages that we and other investigators have found to influence men's earnings. This suggests that parents may influence children's preferences for cash versus psychic income.

REFERENCE

D. Featherman and R. Hauser, "Sexual Inequalities and Socio-economic Achievement in the U.S., 1962–1973," Institute for Research on Poverty, disc. pap. 275–75, Univ. of Wisconsin 1975.

# Intergenerational Transmission of Income and Wealth

By Jere Behrman and Paul Taubman*

Parents can transmit income and wealth to their offspring directly through bequests and gifts of financial assets. Parents can augment a child's skills and traits through the provision of, for example, education, diet and doctors. Biological parents also supply the genetic endowments which provide skills useful in the labor market. Although practically a virgin territory, it is important to estimate the effects of genetic endowments and common (mostly family) environment on earnings and on other indices of individual attainment in order to gauge the effectiveness of programs designed to improve equality of opportunity. Also if genetics and common environment are not controlled for, biased estimates of regression coefficients of observable variables, such as years of education, often are obtained.

In this paper we first summarize some evidence relating to the relative contribution of genetics, common environment, and other environment (as well as covariance terms) to the total variance of four indicators of individual attainment: years of education ($ED$), socioeconomic status of initial civilian occupation ($SES_0$), socioeconomic status of occupation in 1967 around age 45, ($SES_{67}$), and earnings in 1973 around age 50 ($Y_{73}$). We then present regression results for these four indicators, in part to illustrate the importance of biases introduced by ignoring genetic and environmental factors.

For both parts of the paper the empirical estimates are based on the National Academy of Science—National Research Council twin sample which is drawn from white male, veteran twins born between 1917 and 1927. The sample consists of about 2000 twin pairs, roughly evenly divided between monozygotic ($MZ$ or identical) and dizygotic ($DZ$ or fraternal) twins (see Taubman for more details). For both parts of the paper a recursive model is assumed among the above four indicators: education affects socioeconomic scores and earnings, initial occupational status affects mature occupational status, and mature occupational status affects subsequent earnings.

## I. Contributions of Genetics and Environments to Total Variance of Four Indicators of Individual Attainment

In addition to the recursive structure described above, this analysis assumes that there are genetic indices ($G$), two environmental indices ($N$) and four random variables ($v_i$) in the system. The random variables are assumed to be uncorrelated across brothers and equations, with direct effects only on their own indicator. We also assume that genetic effects are additive, that mating is random, that the noncommon environment of one $DZ$ brother is not correlated with his brothers' genes, and that the correla-

* Professors of Economics, University of Pennsylvania. We thank Barbara Atrostic and Shahrokk Fardoust for their able research assistance. This research was supported by National Science Foundation Grant SOC 73-05543.

TABLE 1—PERCENTAGE OF TOTAL VARIANCE ACCOUNTED FOR BY GENETICS AND ENVIRONMENT

| Variable | $\rho$ | Genetics | Total Environment | $G$ and $N$ Covariance | Common Environment |
|---|---|---|---|---|---|
| Education | .65 | 42% | 58% | 0% | 37% |
| | .62 | 28 | 53 | 10 | 33 |
| | .60 | 13 | 50 | 19 | 30 |
| Initial occupation *SES* | .11 | 13% | 86% | 1% | 9% |
| | .13 | 8 | 88 | 2 | 11 |
| | .16 | 4 | 91 | 4 | 15 |
| Later occupation *SES* | .00 | 50% | 50% | 0% | 0% |
| | .10 | 37 | 55 | 4 | 6 |
| | .24 | 4 | 66 | 15 | 16 |
| Ln 1973 earnings | .11 | 52% | 48% | 0 | 5% |
| | .20 | 38 | 53 | 5 | 11 |
| | .30 | 5 | 61 | 17 | 18 |

*Note:* The first entry in each cell is at the maximum genetic contribution, the next at the approximate midpoint of the $\rho$ estimates, and the bottom one at the minimum genetic contribution. $\rho$ is the correlation in the brothers' environment.

tion between brothers' environment is the same for *MZ* and *DZ* pairs. If this correlation is greater for *MZ* than *DZ* pairs, our estimates of genetic effects are upper bounds.

Although the *G*'s and *N*'s are unobservable, we can estimate their variances and covariances and obtain consistent estimates of the coefficients of the measured variables by using data on twins (for details see Behrman and Taubman and Taubman). Each indicator can be expressed in a reduced form containing only *G*'s, *N*'s, $v_i$'s and structural coefficients. Furthermore, in each reduced form, we can collapse the *G*'s into one weighted genetic index and the *N*'s and $v_i$'s into one weighted environmental index. From any one reduced form, we can calculate *ranges* for the proportions of total variance of each indicator accounted for by genetics, environment and their covariance, the correlations in the brothers environment ($\rho$), and the proportion of the variance accounted for by common environments (see Taubman for details). The proportion accounted for by noncommon environment is identified.[1]

Table 1 presents these ranges. In general, the variance of total and common environments are restricted to narrow ranges while genetics and the covariance term cover wide bounds, almost including 0. However, to allow us to summarize this material briefly, we only comment on the midpoint results. At the midpoints, environment by itself accounts for roughly 90 percent of the variance in *SES* in initial occupation and 50 percent of the variance of everything else. Common environment is most important for education at 36 percent, while at 11 percent or less for the other three variables. Since much education occurred when the brothers generally had a relatively large proportion of the environment in common (e.g., living in the same household), the relatively great impact of common environment on

[1] The proportion of the variance accounted for by noncommon environment is the variance due to total environment minus that due to common environment, and is easily calculated from the table.

TABLE 2—BETWEEN AND WITHIN SEMILOG PAIR EQUATIONS[a]

| Dependent Variable: Natural Logarithm of | Between or Average | | | | | | | | Within or Difference | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $SES_F$ | $ED$ | $ED1$ $DUM$ | $ED2$ $DUM$ | $SES_o$ | $SES_{67}$ | $CONST$ | $R^2$ | $\Delta ED$ | $ED1$ $DUM$ | $ED2$ $DUM$ | $\Delta SES_o$ | $\Delta SES_{67}$ | $CONST$ | $R^2$ |
| *MZ* Pairs | | | | | | | | | | | | | | | |
| 1 $ED$ | .0032 | | | | | | 2.48 | .069 | Not estimated | | | | | | |
| | (8.9) | | | | | | (22.59) | | | | | | | | |
| 2 $SES_o$ | .0021 | .090 | .037 | .138 | | | −.149 | .253 | .041 | .058 | −.094 | | | −.020 | .017 |
| | (2.5) | (7.0) | (0.6) | (2.1) | | | (1.0) | | (2.6) | (0.6) | (1.0) | | | (1.0) | |
| 3 $SES_{67}$ | −.0017 | .064 | −.029 | .034 | .063 | | 2.83 | .326 | .054 | .145 | −.179 | .026 | | .035 | .072 |
| | (2.9) | (7.2) | (0.6) | (0.8) | (8.4) | | (28.3) | | (4.4) | (1.9) | (2.4) | (3.1) | | (2.2) | |
| 4 $Y_{73}$ | .0008 | .057 | .165 | .235 | | .0032 | 8.68 | .300 | .010 | .254 | −.227 | | .0024 | .001 | .038 |
| | (1.5) | (9.7) | (2.5) | (3.8) | | (4.2) | (134.4) | | (1.3) | (3.0) | (2.9) | | (3.8) | (0.3) | |
| *DZ* Pairs | | | | | | | | | | | | | | | |
| 1 $ED$ | .0042 | | | | | | 2.43 | .088 | Not estimated | | | | | | |
| | (9.4) | | | | | | (187.9) | | | | | | | | |
| 2 $SES_o$ | .0024 | .091 | .115 | .119 | | | −.213 | .298 | .049 | .181 | −.216 | | | −.0087 | .056 |
| | (2.8) | (7.6) | (2.1) | (2.1) | | | (1.6) | | (3.4) | (1.9) | (2.2) | | | (0.4) | |
| 3 $SES_{67}$ | −.0007 | .049 | .070 | .0058 | .048 | | 3.05 | .285 | .036 | .262 | −.216 | .012 | | −.014 | .091 |
| | (1.3) | (6.0) | (1.9) | (0.2) | (6.1) | | (33.8) | | (3.3) | (3.5) | (2.9) | (1.4) | | (0.8) | |
| 4 $Y_{73}$ | .0021 | .054 | .200 | .180 | | .0036 | 8.64 | .309 | .040 | .293 | −.323 | | .0025 | .0019 | .096 |
| | (3.5) | (8.6) | (3.3) | (3.1) | | (4.1) | (123.0) | | (5.0) | (2.8) | (3.3) | | (3.2) | (0.3) | |

[a] *ED* = years of education; *SES* = socioeconomic score census code; *Y* = annual earnings; *ED1 DUM* and *ED2 DUM* are dummy variables for more than 16 years of education (19 years in the earnings equation) for the first and second brothers, respectively. The subscripts are *F* = father's, *o* = original or initial civilian occupation, 67 = 1967 (or 1972) occupation, and 73 = 1973. The absolute values of *t* statistics are in parentheses beneath the point estimates.

education seems plausible. Genetics by itself accounts for roughly 30 to 40 percent of everything except initial occupation, where it accounts for 8 percent. This last pattern suggests that initial job determination is subject to considerable uncertainty, but that observations over time (including responses to on-the-job training) improve the identification of genetically based characteristics.

If we split up the covariance term proportionately to the variances, we find that *genetics plus common environment* account for 79 percent of the variance in education, 20 percent of the variance in *SES* in initial occupation, 45 percent of the variance in *SES* in later occupation and 53 percent of the variance in subsequent earnings. Alternatively these estimates imply the noncommon environment including measurement error accounts for 21 percent of the variance in education, 77 percent of the variance in initial occupational status, 50 percent of the variance in mature occupational status, and 43 percent of the variance in subsequent earnings.

## II. Regression Results for the Four Indicators of Individual Attainment

Table 2 contains simple semilog[2] regression results for the recursive model described in the introduction. Because of apparent nonlinearities, the years of education variable is supplemented by dummy variables with a value of one if the first or second brother has more than 16 years (19 years in the earnings equations) of formal education. In addition to the observable variables discussed above, these regressions include the socioeconomic status of the fathers ($SES_F$).[3]

This table contains separate panels for *MZ* and *DZ* twins, with equations estimated both for within and between pairs. The within pair results hold constant

[2] The semilog form is presented here because it commonly is utilized in the literature and, in this sample, generally dominates both the nonlog and log-log forms in terms of the degree of consistency with the variance in the dependent variables and in terms of not implying significantly nonzero interaction terms (see Taubman for a test of the latter).

[3] Space precludes the discussion in this paper of other variables considered (e.g., age, hours worked) except to note that their exclusion does not significantly alter the estimates included in Table 2.

genetic endowments and common environment for the *MZ*'s and common environment for the *DZ*'s. The between equations only control for these factors with fathers' *SES*. Comparisons of the within and between results indicate the extent to which there are genetic and common environmental factors that determine the indicators of individual attainment, that are correlated with the right-hand side variables, and that are not adequately proxied by fathers' *SES*.[4]

The between equations suggest that the recursive model in which early occupational attainment affects later occupation which affects earnings is valid. The corresponding *MZ* and *DZ* between equations are similar, as should be expected.[5] There is, however, a tendency for fathers' *SES* to have algebraically large coefficients for *DZ* pairs. Since *DZ* pairs on average come from lower *SES* families, this may indicate that the effect of our *SES* index is not linear. In any case fathers' *SES* tends to make a contribution even when education and other variables are included.[6]

The within equations yield qualitatively similar, but quantitatively disimilar results.[7] Comparison of the between and within and *MZ* and *DZ* estimates suggest that (1) the returns to education in terms of initial occupational status are substantially overstated if common environment is not controlled for and are somewhat overstated if genetics are not controlled for; (2) the returns to education in terms of later occupational status are somewhat overstated if common environment is not controlled for;[8] and (3) the returns to education in terms of earnings are somewhat overstated if common environment is not controlled for and are substantially overstated if genetics are not controlled for.

## III. Concluding Remarks

Several qualifications to this study need be made. (1) The results given in Section I assume random mating, additive genetic effects, independence of the noncommon environment of one *DZ* brother from his sib's genes, and identical correlations between brothers' environments for both types of twins. (2) The results need not generalize to the population as a whole because the distribution of genetics and/or environment may vary for other cohorts. (3) The sample eligibility requirements and differential response rates may have caused the sample to have different distributions of $G$ and $N$ than in the population, although it is not obvious in which dimension the sample is more atypical. We are conducting future work to explore the importance of some of these qualifications.

Subject to the caveat that such explorations do not lead to significantly different results, however, some important conclusions emerge. The results in Section II suggest that the contributions of education, traditionally thought to be one of the most important policy tools in our society for assuring equality of opportunity, are significantly overstated by most estimates in regard to both occupational

[4] There can be other genetic and environmental variables that determine the indicators of attainment, but are not correlated with the right-hand side variables. Also the implied biases are overstated since the variance of measurement error is almost surely a larger portion of total variance in the within variables.

[5] Statistical testing of their similarity is now underway.

[6] The direct effect of $SES_F$ on $SES_{67}$ is negative for both *MZ* and *DZ* twins, but the reduced form indicates that the total effect in both cases is positive.

[7] As previously mentioned, the measurement error problem is probably more severe for the *MZ* equations. However, extremely large measurement errors would be needed to explain the results.

[8] For this indicator the net effects of controlling for genetics in addition to common environment depends on the number of years of education because the coefficient of years of education is larger for the *MZ* cases, but the nonlinear terms are larger for the *DZ* estimates.

status and earnings. The results in Section I, moreover, imply that genetics and common environment (in substantial part, in the home) determine almost four-fifths of the variance in years of education. Thus, to the extent that education does have an impact on occupational and earning mobility, for the most part the degree of education is determined by family characteristics, both genetic and environmental. These family characteristics also have very substantial impact on the variance in mature occupational status and earnings.

Subject to the same caveat, this study suggests to us that an economic-political system that is basically free enterprise will be one in which economic inequality will be passed on from one generation to another via genetic endowments and family environment. Moreover we suspect that the common environment effect includes much more than differential family access to financial markets. Thus policies that go no further than the elimination of money market imperfections will not substantially reduce inequality within or between generations. Of course one might be less concerned about our findings if $G$ and $N$ are distributed independently but we believe that they are positively correlated if for no other reason than family income is correlated with diet and material goods used within the family.[9]

[9] Paradoxically allowing for a positive $G$ and $N$ covariance reduces the genetic contribution (even after apportioning the covariance).

REFERENCES

**J. Behrman and P. Taubman,** "Earnings, Schooling, and Ability: A Multiple Indicator Model Using Twins," Univ. of Pennsylvania, Philadelphia, mimeo, 1975 (Presented at Toronto World Econometric Congress, 1975).

**P. Taubman,** "The Determinants of Earnings: Genetics, Family and Other Environments: A Study of White Male Twins," Univ. of Pennsylvania, Philadelphia, mimeo, 1975 (Presented at San Francisco Econometric Society Meetings, 1974).

# EXCHANGE VERSUS GRANT TRANSACTIONS IN ENVIRONMENTAL MODELS

# Incentive for Innovation as the Basis for Effluent Charge Strategy

*By* LLOYD ORR*

As a basis for sound environmental policy, most economists favor effluent charges or some similar market surrogate. This appears to stem from the common view of prices as effective communication and incentive mechanisms and the substantial difficulties—experienced and foreseen—of implementing efficient regulatory schemes over space, over different classes of pollutants, and over time. However, effluent charges have been subjected to challenge and debate. Questions have been raised concerning both theoretical efficiency and practical problems of implementation.

Ronald H. Coase has argued that where negotiation is costless the *efficiency* of the result is not dependent on the assignment of property rights to those damaged by an externality. He makes the further argument that when negotiation is not possible the imposition of taxes, or taxes and subsidies respectively on the producers and consumers of a damaging externality, may not be the most efficient policy. This conclusion appears to stem from two related considerations: (1) it may be cheaper for the consumer to protect himself from the damage than for the producer to eliminate it, and (2) the extent of the damage is dependent on consumer as well as producer behavior. Laundries affect the amount of damage from factory smoke by their location decisions and, indeed, in the presence of a subsidy will tend to locate near the factory in greater than optimal numbers.

Several authors have expanded on this theme and elaborated on implications of group size and the presence of some form of market power. Conclusions tend toward the view that simple tax or tax-subsidy schemes could not be expected to yield efficient solutions to externality problems. In a synthesis of this literature Ralph Turvey concludes that when negotiation is not possible, "the theorist should be silent and call in the applied economist"—suggesting that no prescriptions of a general nature are possible.

Economists who are accustomed to thinking that allocative efficiency is associated with the payment of marginal cost for use of a factor may well find themselves uncomfortable with this particular branch of the literature on externalities. Indeed, William Baumol has provided a substantial basis for skepticism. Using a simple general equilibrium model he demonstrates that the optimal price for an externality-generating product is its marginal private cost plus its marginal external cost. The correct prescription is a "Pigouvian" tax on the externality without com-

* Associate Professor of Economics, Indiana University.

pensation to those damaged. His interpretation is a simple and clear statement of the public goods problem applied to negative externalities. Just as efficiency requires the producer of a positive externality to be compensated at price equal to marginal cost, the producer of a negative externality should be similarly charged. Also, since the consumer of a positive "public" externality adds nothing to cost, the appropriate price is zero, and similarly since the consumer of a negative public externality detracts nothing from the damage done to others, his appropriate compensation is also zero. Lump sum transfers are, of course, excepted from this prescription.

Baumol attributes the difference in his conclusion from that of Coase to the existence of local maxima in the Coase example that would prevent achievement of the global maximum through use of a Pigouvian tax. He also attributes some difficulties in the Coase argument to confusion between technological and pecuniary externalities.

Given the inordinate amount of information needed to identify the global maxima required for optimal effluent charges and related problems such as the lack of perfect competition, many writers have fallen back on the position of using effluent charges, established through an iterative process, as an effective means of reaching preestablished standards of environmental quality. Baumol and Wallace E. Oates have shown that such a strategy will achieve the standard at minimum cost provided only that firms produce whatever output they select at minimum cost.

There are, however, many practical problems associated with the implementation of these effluent charge structures. To mention a few, there are monitoring difficulties, problems of geographical differences in charge structures, and the possibility of excessive adjustment costs associated with the iterative search for appropriate charge levels. Anthony Dorcey has noted that the iterative process could lead to the very sort of protracted bargaining and regulatory problems from which effluent charges are supposed to provide an escape.

However the basic difficulty that I have with the effluent charge literature is quite different. It is the tendency to concentrate excessively on short run problems of allocative efficiency. It seems to me that the greatest advantage of effluent charges relative to alternative control mechanisms is in their provision of decentralized incentives for technological change. It also strikes me that it is most appropriate to view the environmental problem as one of dynamic adjustment to increasing relative scarcity of a resource. Technological adaptation rather than resource allocation becomes the key to an effective solution.

Technological change is nearly always mentioned in discussions of effluent charge strategies, and it occasionally (see Richard O. Zerbe) becomes an integral part of the analysis. What is missing is the view that environmental policy is fundamentally the need to establish a framework that provides *continuous* and *detailed* technological adaptation to the impacts on the environment of growth, change in product mix, and change in process technology. It is in this context that effluent charges and other strategies should be evaluated—both in terms of effectiveness and political feasibility. In essence, emphasis shifts from welfare economics to growth with exhaustible resources.

## I. Technical Change with Exhaustible Resources

One positive impact of the recent doomsday models is that they have stimulated some good work in the economics of "exhaustible" resources. Of particular interest here is the literature on growth paths with

exhaustible resources.

Although the notion of an exhaustible resource seems intuitively clear, it turns out to be a bit slippery when you try to tie it down. A definition might be the following: given current prices and technology, if the rate of use of a resource stock exceeds its rate of renewal, then the resource is exhaustible. P. Dasgupta and Geoffrey M. Heal note that this creates a "problem" only under the further condition that the resource is "essential"—that is, production declines to zero in its absence.

It is easy to place environmental quality under this concept of exhaustible and essential resources, and thus to assert the relevance of recent growth models for environmental analysis. Deterioration in environmental quality signals that the "stock" of resources is being depleted and thus that use exceeds renewal. The necessity of survival for a wide variety of biological species signals the essentiality of determinable minima in at least some dimensions of environmental quality.

My interest in the growth models with exhaustible resources lies in the focus they provide on the avenues of escape from resource limitations (Robert M. Solow). As one might intuit, escape depends crucially on the elasticity of substitution between renewable and exhaustible resources and on technical progress that reduces the exhaustible resource input per unit of output. Doomsday occurs if a resource remains in the essential category and if there is some upper bound on real output per unit of the resource that is not too far from our present position.

Economic historians provide us with valuable additional focus on these matters.[1] In their traditional concerns with technical progress as a response to relative factor scarcities and with the question of factor "bias" in innovation, the analytical distinction between technological change and factor substitution is extremely difficult to maintain. In Rosenberg's words (1973), "*Today's* factor substitution possibilities are made possible by *yesterday's* technological innovations." We thus see the problem of ease in substitution as essentially one of innovation in creating substitution possibilities, and the problem of relief from future resource scarcities comes to rest squarely on our ability to induce technological response.

There is another dimension in which historians have something important to say about our response to future resource scarcities. In their assessment of the importance of technological change in past growth, they have provided valuable documentation (1) on the vast range of technological response possibilities to particular resource scarcities, and (2) on the *detailed adaptation to such scarcities over time under the spur of decentralized cost incentives.*[2]

It is not possible to demonstrate with certainty that allocative efficiency is less important than technological innovation in solving a particular resource problem at a particular time—especially since the two are not independent. However, where the problem is one of continuous and long run adjustment to growing scarcity of an essential resource, as it clearly is in the case of environmental quality, a very strong case can be made for a shift in emphasis in policy formation from allocative efficiency to long-run technical adaptation.

## II. Experience with Effluent Charges

Although data on experience with effluent charges are only beginning to accumulate, there have been a variety of reports as to their effectiveness. These reports have usually been in connection with muni-

[1] See Nathan Rosenberg (1972, 1973) and Paul A. David, pp. 57–91.

[2] Rosenberg (1973), G. Hueckel.

cipal sewer charges and in-plant water and waste charges.

Reports on the experience of individual firms or municipalities frequently show dramatic impacts from the imposition of effluent charges.[3] Response begins soon after imposition and effectiveness increases over time. Reductions of 60 percent or more in biochemical oxygen demand (*BOD*) load occur over periods of one to three years. Technological response by individual firms frequently results in *favorable* effects on cost that are in addition to the reduction in sewage charges.

Ralph D. Elliott and Seagraves have studied the impact of surcharges on industrial waste using a pooled time series and cross-section sample of 198 observations from 34 cities. Although there was a great deal of "noise" in the data, the negative price-quantity relationship showed a substantial impact following the imposition of moderate waste charges. In an interpretive statement it is suggested that a 45 percent reduction in *BOD* demand would follow from the imposition of a "modest" surcharge of 2.7 cents per pound of *BOD*.[4]

Some caution needs to be used in interpreting these reports. The statistical results are less conclusive than one might wish, and the case study selections may be somewhat biased toward the dramatic. Also the in-plant charges and sewage surcharges do not represent comprehensive effluent charge schemes. In general we do not know what happened to the residuals that are no longer dumped into the sewage system. If there were adequate charges covering atmospheric and solid waste disposal and direct release of treated or untreated wastes into water courses, the results might not have been so dramatic.

However, the reports do illustrate some characteristics of the cost incentives that are inherent in effluent charges. The setting of these charges close to social optima would be a chance occurrence, since information for such charge levels was generally lacking. Nevertheless, we see managers responding to resource cost in a variety of ways that are suited to their own particular circumstances. The nature of the response would be difficult to predict—either for someone trying to plan a specific result, or even for the waste dischargers themselves. The reports strengthen the notion that environment-saving technology is a sparsely explored area that yields quickly to the proper set of incentives. The result is that charges have been effective, their effectiveness has increased with time, and they apparently have not been excessively disruptive. The strength of the partially qualitative and less than conclusive nature of these results is enhanced by the strongly negative assessments of current environmental regulations with respect to incentive structures and technological change.[5]

### III. Effluent Charge Policy

One might well ask why the distinction between allocative and innovative impacts of effluent charges is deserving of emphasis. Clearly, effluent charges designed to meet allocative criteria would, in their reflection of relative scarcities, create incentives for innovation in the proper directions. My response is twofold. First, the importance of innovation makes the extent of our ignorance as to proper price structures more clearly recognizable. The required information for establishing

[3] James A. Seagraves, p. 875; Allen V. Kneese and Blair T. Bower, pp. 168, 170–172; Hearings, testimony by Robert H. Haveman, L. Kimball, and J. Kinney.

[4] Pages 3, 27–28. It is noted that this level can only be achieved after an elapse of time required for adaptation. It is suggested (p. 6) that case studies of industrial groups subjected to effluent charges for several years "should permit description of major adjustments in industrial processes."

[5] A. Myrick Freeman and Haveman.

theoretically correct price structures is not merely inordinately difficult to acquire—it is nonexistent. Second, emphasis on the need for thorough, detailed, and continuous adaptation to the social costs of using common property resources highlights the most compelling advantage of effluent charges in comparison with alternative environmental policies. Taken together these two factors bring about a shift in the criteria for a "good" effluent charge strategy. The result should be a reduction in both the political and economic problems of implementing an effective policy.

Recognition of the difference between allocative and innovative price structures is important. If we could establish a set of Pareto-optimal prices, or even a time pattern of prices designed to maintain given effluent standards, they would not in general be the appropriate prices for optimal innovation. Optimal innovation prices presumably would be those that induce innovative activity that maximizes discounted net returns. Quite apart from the intractable problems of selecting an appropriate social discount rate, the parameters for the production function of knowledge cannot be known.[6] Therefore, the data necessary for calculating the "correct" price cannot be generated.

Our inability to "fine tune" environmental quality should be obvious regardless of the policy approach. But, awareness of the limits of our knowledge should not lead to paralysis or the abandonment of environmental policy to traditional styles of regulation. It rather suggests gradualism as a means of minimizing the uncertainties of transition cost and flexibility in response mechanisms as a means of inducing effective use of current and prospective knowledge and as a guard against massive changes in a complex milieu where side effects are largely unknown.[7] Effluent charges become very imperfect tools, but, in my view, clearly superior to known alternatives as a *basic* policy.

Emphasis on establishing a long run framework of technological adaptation effectively reduces some of the economic and political problems associated with the implementation of effluent charge strategies. For example, many of the economic and political problems of effluent charges are related to transition costs. Gradual implementation of charges in predetermined steps over a period of years should provide the needed inducement for innovation while minimizing short term disruptions. Rather than being faced with highly uncertain levels of effluent charges, the waste discharger responds to a fairly stable profile of charges extending several years into the future. Environmental quality standards may still be relevant, but the *direction* and *rate of change* in environmental quality are criteria that should be of more continuing importance.

As the above implies, there can be a marked reduction in the current stress on cumbersome and potentially costly iterative adjustments in charge structures. Proper charge levels remain an important question, and there are substantial information requirements for establishing a desirable *ordering* of charges for the major pollutants and levels that will induce an initial response. The reported experiences we have noted suggest that this may be surprisingly easy. The iterative process would still be useful where

[6] Rosenberg (1972, p. 172).

[7] C. Holling and Michael A. Goldberg note the similarities between complex economic (urban) and ecological systems. Both appear to have considerable stability within certain bounds, but, when the complex balance is disturbed by intervention, they can generate unexpected and undesirable results. Policy actions that are limited in scope and diverse in nature are suggested, using principles based more on recognition of ignorance than the presumption of knowledge about the nature of the system.

initial charge structures fail to elicit adequate response or where the burdens are clearly onerous and not justified by actual or anticipated response. Errors will still need to be corrected, but the abandonment of fine tuning ambitions carries with it a marked decrease in the use of an iterative process.

Another implication of emphasis on innovative response is a decrease in the importance of regional differences in charge levels. The possibility that regional differences in charges will seriously affect the competitive position of plants *within* an industry at different locations is likely to make substantial differentials politically unacceptable. Also, in the absence of knowledge on technological response to a basic nationwide charge level, we cannot know that the gain in allocative efficiency from differential charges, based on current parameters, would be worth the costs of disruption and relocation. Thus for both political and economic reasons it seems wise to start with a gradual implementation of nationwide charges. If the short-run transition burdens from the charges are still regarded as too heavy, other business taxes could be reduced. This would provide general relief without destroying the incentive to reduce effluent or the valuable differentiation (allocative and innovational) *between* industries on the basis of quantity and quality of waste discharge. Local governments with special problems may want to provide supplementary charges or regulation. This is clearly appropriate, especially since they are likely to be more sensitive to the consequences of their actions than the federal government.

Ignorance is not a good thing, but its *recognition* can be very valuable in helping us to avoid serious errors. The strength of effluent charges lies in their ability to move things in generally desirable directions even when we lack the knowledge to provide price structures that are efficient in either the allocative or innovative sense. The diversity of impact, continuous incentive properties, and relatively high degree of self regulation serve to limit the undesirable (and unforeseen) side effects and the perverse incentives that are associated with many regulatory policies. The relevant concept is *robustness*—the ability to function effectively in the face of the often serious differences between the ideal of the model and the reality of operative policy. This concept is relevant to all policy formulation and its serious consideration should, it seems to me, prevent some of the disaffection with the government sector that follows from recurrent policy failures.

## REFERENCES

W. J. Baumol, "On Taxation and the Control of Externalities," *Amer. Econ. Rev.*, June 1972, *62*, 307–22.

——— and W. E. Oates, "The Use of Standards and Pricing for Protection of the Environment," *Swedish J. Econ.*, Mar. 1971, *73*, 42–54.

R. H. Coase, "The Problem of Social Cost," *J. Law Econ.*, Oct. 1960, *3*, 1–44.

P. Dasgupta and G. M. Heal, "The Optimal Depletion of Exhaustible Resources," *Rev. Econ. Stud.*, 1974, *Symposium*, 3–28.

P. A. David, *Technical Choice, Innovation and Economic Growth*, London and New York 1975.

A. Dorcey, "Effluent Charges, Information Generation, and Bargaining Behavior," *Nat. Res. J.*, Jan. 1973, *13*, 119–33.

R. D. Elliot and J. A. Seagraves, *The Effects of Sewer Surcharges on the Level of Industrial Wastes and the Use of Water by Industry*, Raleigh, Aug. 1972.

A. M. Freeman III and R. H. Haveman, "Clean Rhetoric and Dirty Water," *Public Interest*, Summer 1972, *28*, 51–65.

C. S. Holling and M. A. Goldberg, "Ecology and Planning," *Am. Inst. of Planners J.*, July 1971, *37*, 221–30.

G. Hueckel, "A Historical Approach to Future

Economic Growth," *Science*, Mar. 1975, *187*, 925–31.

A. Kneese and B. Bower, *Managing Water Quality: Economics, Technology, Institutions*, Baltimore 1968.

N. Rosenberg, "Innovative Response to Materials Shortages," *Amer. Econ. Rev. Proc.*, May 1973, *63*, 111–18.

———, *Technology and American Economic Growth*, New York 1972.

J. A. Seagraves, "Industrial Waste Charges," *J. Environ. Engineering Div., Amer. Soc. Civil Engineers*, Dec. 1973, *99*, 873–81.

R. M. Solow, "The Economics of Resources or the Resources of Economics," *Amer. Econ. Rev. Proc.*, May 1974, *64*, 1–14.

R. Turvey, "On Divergences Between Social Cost and Private Cost," *Economica*, Aug. 1963, *30*, 309–13.

R. O. Zerbe, "Theoretical Efficiency in Pollution Control," *West. Econ. J.*, Dec. 1970, *8*, 364–76.

U.S. Congress, Joint Economic Committee, Hearings before the Subcommittee on Priorities and Economy in Government, *Economic Analysis and Efficiency of Government: Part 6, Economic Incentives to Control Pollution*, Washington 1971.

# Federal Environmental Policy and R&D on Water Pollution Abatement

*By* HAROLD J. BRUMM, JR. AND DANIEL T. DICK*

Keynes once asserted that the current economic cant of politicians represents the doctrine of some academic scribbler of an earlier day. We suspect he would have been tempted to disavow ever having made that claim if he had had the opportunity to witness the development of federal water pollution control policy over the past two decades, because it is painfully apparent that the writings of economists (past and present) have had precious little impact on the formulation of that policy (Edwin S. Mills and Frederick M. Peterson). The four central components of federal water pollution control policy are: (1) prohibition of discharge of any refuse into navigable waters without a permit from the Army Corps of Engineers; (2) subsidies of up to 75 percent of the construction costs of municipal water treatment facilities (*WTF*'s); (3) ambient standards (e.g., the amount of dissolved oxygen in parts per million) on all navigable waters in the United States; and (4) a five-year tax amortization of additions to industrial waste treatment plants.[1] This grants and regulations strategy has been thoroughly scored by several economists, who have urged instead that greater reliance be placed upon effluent charges, i.e., taxes levied on pollutants discharged into watercourses (see, e.g., Kneese and Schultze, William Baumol and Wallace Oates, and A. Myrick Freeman).

The twofold purpose of this paper is to briefly review the environmental policy recommendations which appear to have widespread support in the economics profession, and to examine their implications for investment in water pollution abatement technology. In addition, we shall proffer a modest policy recommendation that we have not encountered anywhere in the literature.

## I. The Justification for Government Intervention

The existence of a federal pollution control policy raises two questions. First, why is government intervention required at all? Several writers (Ronald H. Coase, James Buchanan and W. Stubblebine, Otto A. Davis and Andrew B. Whinston, Harold Demsetz, and Dean Worcester to name a few) have argued that government intervention may not only be unnecessary, but ineffectual at best—particularly if the policy is based on a Pigouvian tax-subsidy scheme. Evidently it is their view that bargaining, which is a form of market exchange solution, holds greater promise for a socially efficient allocation of resources. However, Dick (forthcoming) shows in a review and synthesis of the literature critical of the voluntary approach that private bargaining is unlikely to yield optimal results re-

* Research Staff Member, Institute for Defense Analyses (*IDA*), and Senior Economist, Resource Development Associates, respectively. The publication of this paper does not indicate endorsement by *IDA*, nor should the contents be construed as reflecting the official position of that agency. Dr. Brumm's research has been supported by a grant from the National Science Foundation *R&D* Assessment Program.

[1] For a discussion of the history of federal air and water pollution legislation, see Allen V. Kneese and Charles L. Schultze.

gardless of liability placement, except in states of nature that are extremely improbable. Any realistic assessment of real world problems in which income effects, transactions costs, and nonseparabilities abound rules out reliance on voluntary modes of control. F. Trenery Dolbear, Ezra J. Mishan, Alan J. Randall, William C. Wheaton, and David K. Whitcomb have provided further sobering thoughts on the efficacy of nongovernmental solutions to problems created by external diseconomies.

Given, then, that the unhappy case is made for governmental solutions being the only viable alternative to internalizing various externalities which create environmental problems, a second question arises. What policy instrument (or combination of instruments) should the government employ, in order to achieve a level of pollution abatement that is socially desirable (however defined)? The various collective strategies available for this purpose include (1) regulation, (2) prohibition, (3) subsidies, and (4) effluent charges (Mishan, pp. 14–16).

## II. An Evaluation of Alternative Government Strategies for Water Pollution Control

The static inefficiencies of direct regulation are well known to economists. For example, a regulation requiring all polluters to make the same percentage reduction in waste discharges violates the *intersource* efficiency condition that private marginal costs of abatement be equalized for all pollution sources (Dick, 1974, pp. 67–70). Moreover, a regulation requiring that certain abatement devices or processes be used does not satisfy the *intrasource* efficiency condition that the polluter's marginal costs of abatement be equalized for all pollution control devices or processes which he chooses (Dick, 1974, pp. 70–72). Finally, direct method regulation can be expected to have perverse incentive effects on the invention and development of new pollution control technologies. If, instead of mandating that only one particular abatement method be employed, the government were to announce that a particular standard would have to be met but the use of a wide range of alternative methods would be allowed, we may anticipate that there would be a greater number of competitors in the market for pollution control *R&D*, thereby enhancing the possibility that truly significant abatement technologies would be developed.

Outright prohibitions of pollution activity are, in general, economically inefficient because they ignore possible tradeoffs between different levels of environmental quality and other valued goods and services. In a wide range of circumstances pollution should be reduced only up to the point where the social cost of additional abatement equals the social benefit derived from the reduction in damages resulting from that additional abatement. However, in certain cases the prohibition strategy may make sense. An example would be a circumstance of extreme uncertainty about irreversible effects of growing concentrations of a given pollutant.

In evaluating subsidies versus charges there exists a potentially serious "moral hazard" problem (Kenneth J. Arrow, p. 142) inherent in a subsidy scheme. Specifically, a subsidy policy may make it profitable for a firm to pollute more than it otherwise would have, in order to qualify for large subsidy payments; conceivably an enterprise which in the absence of a subsidy scheme would not be economically viable might find the waste-making business profitable (M. Kamien, Nancy Schwartz, and Dolbear; and D. Bramhall and Mills). Another perverse incentive of a subsidy scheme is the in-

ducement it provides for overinvestment in *WTF*s relative to other modes of pollution control. Since up to 75 percent of the construction costs of *WTF*s is borne by the federal government, purchasers of *WTF*s have an incentive to meet standards based on their perceived capital costs of *WTF*s—costs which are net of the federal subsidy. Another objection to a grants scheme is that the change in perceived capital costs it induces invites a capital-intensive bias to production in subsidized firms and municipalities. As a result, waste reduction will be accomplished at the expense of alternative input combinations which may represent a more efficient allocation of society's scarce resources. (The current federal policy of extending the tax advantage of accelerated depreciation to additions to industrial *WTF*s is, of course, a particular type of subsidy; as such, it is susceptible to the foregoing criticisms.)

In summary, one principal liability for the long term of a strategy utilizing prohibition, regulation, or subsidies is that such an approach does not impact on private economic incentives in a way that would encourage the discovery and adoption of new and better technologies for improving water quality. By contrast, as Kneese and B. Bower (Chs. 6 and 7) have pointed out, a water management system relying heavily on effluent charges not only would tend to adjust the degree of pollution reduction required of each activity so as to minimize the total cost of cleanup of the watercourse, but would exert a continuing incentive on polluters to adopt new technologies, as they become available, which are designed to achieve still further reductions in discharges of waste residuals.

On the basis of the foregoing theoretical arguments, it would appear that effluent charges are superior to other water pollution control policy instruments. However, Baumol and David Bradford have alleged that detrimental externalities of sufficient strength will result in a breakdown in the convexity of the social production set. These externality-induced nonconvexities would seem to rob the pricing mechanism of its normative usefulness. Advocates of Pigouvian taxes can take comfort, therefore, from Vernon K. Smith's analysis which demonstrates that the imposition of effluent charges can induce technological changes which serve to ameliorate those nonconvexities, and from Jacob Schmookler's findings that economic incentives have a positive influence on technological innovation. Accordingly, it seems clear that one of the principal advantages of a strategy which relies heavily on effluent charges is its implications for advancement in water pollution abatement technology. Kneese and Schultze have summarized the point succinctly: "Over the long haul, perhaps the most important single criterion on which to judge environmental policies is the extent to which they spur new technology toward the efficient conservation of environmental quality" (p. 82).

Finally, it should be noted that the distributional consequences of subsidies and effluent charges differ. Government expenditures to make subsidy payments must be financed by withdrawing resources from the private sector through increased taxes or borrowing, if the level of nonenvironmental publicly provided goods and services is to remain constant. On the other hand, if effluent charges are employed, the initial burden of the cost falls on the polluters. However, the ultimate incidence of the pollution abatement cost under either an effluent charge or a subsidy policy cannot be determined a priori. In this regard it is worth mentioning that Thomas Havrilesky (p. 327) has argued that the burden of pollution abatement, as a percentage of income, is

borne most heavily by the poor, under a subsidy scheme (or under regulation).

Even if one accepts Smith's argument that the difficulties caused by externality-induced nonconvexities can be overcome, there remain other potentially serious drawbacks to the implementation of effluent charges. For example, Kneese and Bower (Ch. 6) have pointed out numerous problems, including the possibility of nonseparabilities: various wastes discharged into a given watercourse may interact chemically, with the result that the external cost of one pollutant cannot be determined unless the level of discharge of the others is known. A related limitation concerns the measurement of environmental quality and extent of culpability for environmental damage. More generally, the informational requirements of a socially optimal level of effluent charges raise implementation costs to very high levels. (The same problem arises with any other policy instrument, however.) Indeed, since actual measurement of social welfare is impossible, and since the pollution abatement level required for the attainment of a social optimum is unknown except at very high costs, one cannot determine if a given variation in effluent charges will move society toward that optimum. (For detailed critique of the effluent charges strategy, see Susan Ackerman.)

In light of the foregoing arguments, it is clear that optimal effluent charges are an extremely expensive way to achieve effective water pollution control. However, just because there exist problems with the application of effluent charges in the real world, it does not necessarily follow that they should be summarily ruled out. A more sensible approach is the Environmental Pricing and Standards (*EPS*) system suggested by Baumol and Oates (Ch. 10). The *EPS* system entails establishing (somehow) an ambient quality standard, setting a tax on effluents discharged, and then raising the tax rate (which induces polluters to curtail their discharges of wastes) until the preestablished standard is met. A problem with this approach is common to the provision of virtually any public good, namely, determining what the socially optimal level of provision is. In this regard, Baumol and Oates (pp. 149–150) have provided some useful insights.

Under the *EPS* system the standard is set (perhaps through the political mechanism) and the effluent charge is successively raised until the standard is met. That way, whatever standard is met, at least it is met in a least cost manner. The informational requirements of the *EPS* approach are relatively small when compared to those required for setting optimal effluent charges, because for the latter scheme one must have considerably more knowledge of both the basin-wide marginal damage function and the basin-wide marginal cost function. This is the tremendous informational cost requirement that led Baumol and Oates to advocate the *EPS* system in the first place.

We propose that an allocative gain could be made by comparison with the *EPS* system at an informational cost that is a fraction of that associated with the optimal effluent charges plan. This gain can be achieved by charging a variable rate per unit of effluent based on the basin-wide marginal external damages function. While knowledge of the basin-wide marginal external damages function is required, this approach does not require knowledge of the basin-wide marginal abatement cost function. Herein lies the informational cost advantage. If the effluent concentration is monitored at frequent intervals and the charge varied accordingly along the marginal damages function, an equilibrium will be achieved automatically. The allocative gain accrues

because, while the basin-wide authority does not know what abatement costs are, the private decision-making units do. If there are private gains to be had through abatement at the current effluent tax, polluters will exploit these gains and the charges will then be adjusted downward. This process will continue until no further private gains can be achieved through extra abatement and therefore the charge will not be further adjusted.

It should be pointed out that for this scheme to work, the monitoring and adjustment of charges must be made at frequent intervals. This is necessary to avoid having irreversible overabatement take place if early charge levels are high, or underabatement if they are low. Whether or not the informational cost of acquiring knowledge of the marginal damages function is worth the allocative gain achieved over an *EPS* system cannot be determined a priori.

### III. Conclusion

We have attempted to briefly summarize the literature addressing alternative policy instruments for water pollution control, and to reaffirm the case for implementing a policy based on effluent charges. Policy options that rely on private bargaining or regulatory devices will be faced with greater problems of information acquisition, transaction costs, and creation of perverse economic incentives.

From a practical standpoint, the *EPS* approach has considerable appeal, in terms of overcoming some of the aforementioned problems. However, if the basin-wide marginal damage function is known, a system of progressive charges based on this function will improve allocation compared to that of the *EPS* system. The costs of this allocative gain are those associated with determining the marginal damage function. This means that if the basin-wide marginal damage function is known, one need not expend resources to determine the basin-wide marginal cost function, in order to use a charges scheme with optimality properties.

### REFERENCES


**L. Abrams and J. Barr**, "Corrective Taxes for Pollution Control: An Application of the EPS System to Agriculture," *J. Envir. Econ. Manage.*, Dec. 1974, *1*, 296–318.

**S. Ackerman**, "Effluent Charges: A Critique," *Can. J. Econ.*, Nov. 1973, *6*, 512–28.

**K. Arrow**, *Essays in the Theory of Risk-bearing*, Chicago 1971.

**W. Baumol**, *Welfare Economics and the Theory of the State*, 2d ed., Cambridge 1965.

——— **and D. Bradford**, "Detrimental Externalities and Non-Convexity of the Production of the Production Set," *Economica*, May 1972, *39*, 160–76.

——— **and W. Oates**, *The Theory of Environmental Policy*, Englewood Cliffs 1975.

**D. Bramhall and E. Mills**, "A Note on the Asymmetry between Fees and Payments," *Water Resources Res.*, 3d Quarter 1966, *2*, 615–16.

**J. Buchanan**, "Joint Supply, Externality, and Optimality," *Economica*, Nov. 1966, *33*, 404–15.

——— **and W. Stubblebine**, "Externality," *Economica*, Nov. 1962, *29*, 371–84.

**R. Coase**, "The Problem of Social Cost," *J. Law Econ.*, Oct. 1960, *3*, 1–44.

**O. Davis and A. Whinston**, "On Externalities, Information, and the Government-Assisted Invisible Hand," *Economica*, Aug. 1966, *33*, 303–18.

**H. Demsetz**, "The Exchange and Enforcement of Property Rights," *J. Law Econ.*, Oct. 1964, *7*, 11–26.

**D. Dick**, *Pollution, Congestion, and Nuisance*, Lexington 1974.

———, "The Voluntary Approach to Externality Problems: A Survey of the Critics," *J. Envir. Econ. Manage.*, (forthcoming).

**F. Dolbear**, "On the Theory of Optimal Externality," *Amer. Econ. Rev.*, Mar. 1967, *57*, 90–103.

**A. M. Freeman**, "Grants and Environmental Policy," in K. Boulding, M. Pfaff, and A.

Pfaff, eds., *Transfers in an Urbanized Economy,* Belmont 1973.

T. Havrilesky, "Technological Innovativeness, the Grants Economy, and the Ecological Crisis," in K. Boulding, M. Pfaff, and A. Pfaff, eds., *Transfers in an Urbanized Economy,* Belmont 1973.

M. Kamien, N. Schwartz and F. Dolbear, "Asymmetry between Bribes and Charges," *Water Resources Res.,* 1st Quarter 1966, *2,* 147–57.

A. Kneese and B. Bower, *Managing Water Quality,* Baltimore 1968.

——— and C. Schultze, *Pollution, Prices, and Public Policy,* Washington 1975.

E. Mills and F. Peterson, "Environmental Quality: The First Five Years," *Amer. Econ. Rev.,* June 1975, *65,* 259–68.

E. Mishan, "The Post-War Literature on Externalities: An Interpretive Essay," *J. Econ. Lit.,* Mar. 1971, *9,* 1–28.

H. Peskin and E. Seskin, eds., *Cost-Benefit Analysis and Water Pollution Policy,* Washington 1975.

A. Pigou, *The Economics of Welfare,* 2d ed., London 1924.

A. Randall, "Market Solutions to Externality Problems: Theory and Practice," *Amer. J. Agr. Econ.,* May 1972, *54,* 175–83.

R. Ridker, *Economic Costs of Air Pollution,* New York 1967.

J. Schmookler, *Invention and Economic Growth,* Cambridge 1966.

V. Smith, "Detrimental Externalities, Nonconvexities and Technical Change," *J. Publ. Econ.,* Aug. 1975, *4,* 289–95.

R. Turvey, "On Divergences between Social Cost and Private Cost," *Economica,* Aug. 1963, *30,* 309–13.

W. Wheaton, "On the Possibility of a Market for Externalities," *J. Polit. Econ.,* Sept.-Oct. 1972, *80,* 1039–44.

D. Whitcomb, *Externalities and Welfare,* New York 1972.

D. Worcester, "A Note on 'The Postwar Literature on Externalities: An Interpretive Essay,' " *J. Econ. Lit.,* Mar. 1972, *10,* 57–59.

# The Anatomy of Nonmarket Failure: An Examination of Environmental Policies

*By* RON D. WHITE*

In a very well-known article Francis M. Bator synthesizes the work of many other economists and develops a structure for discussing the phenomena of market failure; he distinguishes five modes of market failure and, subsequently, identifies three causes for these failures. Rather than "solving" the problem of market failure, Bator gives us a new set of perspectives from which to study the situation. There is a parallel need to identify and distinguish various modes and causes of failure in our nonmarket mechanisms. Public policy analysts need to make an effort to generalize about the circumstances under which public policies fail. Although many analysts have discussed various aspects of institutional failure, there has been no attempt to organize the topic and provide a typology from which further analysis can proceed. This paper first discusses problems encountered in defining nonmarket failure, then develops a classification system for such failures, and applies this framework to environmental policy.

Academic work has gone on for years to describe the circumstances in which markets fail in the distribution of goods and incomes. Most public finance texts spend considerable time in developing these models. By comparison, there is almost no analysis of the circumstances in which public programs and policies have failed. Well-organized assessments of the performance of publicly supplied services and/or publicly financed programs, which one would expect to find, are usually missing from the texts.

It is hardly the case that analysts have failed to find fault with various public programs. Rather, the problem is that these criticisms have never been systematically synthesized and used as an alternative to the "market failure" approach. The market failure literature is informative but, because it analyzes markets, it is of limited usefulness as a basis for public policy. The most important question for policy makers and policy analysts is: How does existing nonmarket decision making compare with an idealized nonmarket system? This query has not been pursued in any systemic fashion. The purpose of this work is not to present answers, but to try to focus attention on the question.

## I. Definitional Problems

Nonmarket failure does not have any agreed upon definition and the term is not one in standard usage. Because public policies are criticized by such a variety of scholars with diverse disciplinary attachments and political perspectives, no widely-used definition has emerged. Nor is there a standard analogous to "market efficiency" against which nonmarket institutions can be judged. In a typical policy-making situation, where redistribution occurs or will occur, there are no purely

* Director, Center for Business and Economic Research, Wichita State University, and The Federal Energy Administration.

economic criteria available. Yet, there may be considerable agreement—although not unanimity—that specific policies have failed. Thus the development of a typology for nonmarket failure is feasible so long as complete agreement is not required in the definition of "nonmarket failure."

On the other hand, market failure has a rather well-accepted definition which is a central part of modern welfare economics and which requires the acceptance of some strong assumptions about taste, technology, and other factors. Nonmarket failure has no similar, formal definition. Bator does provide a basis from which to proceed, however. His first definition of market failure is: ". . . the failure of a more or less idealized system of price-market institutions to sustain 'desirable' activities or to estop 'undesirable' activities" (p. 351). A parallel definition of nonmarket failure can be stated. Nonmarket failure is a breakdown of the system of taxation, transfers, and expenditures which results in the lack of progress toward stated goals or the active pursuit of unsanctioned aims. The system of taxation, transfers, and expenditures is, of course, the stuff of empirical public finance. Yet the usual textbook approaches these questions from the market-system perspective.

The nonmarket conceptualization of these processes is quite different. Kenneth E. Boulding's classification of "exchange relationships" and "threat-submission" relationships and "integrative" relationships is very useful here. Exchange is the basic market relationship in which participants mutually agree to trade resources. Threat-submission relationships involve the extraction of resources by use of negative sanctions. Integrative situations are those in which resources are distributed or redistributed on the basis of status. The market, or exchange, approach to public policy has been shown to be unsatisfactory by formal analysis and—it is contended here—gives an unduly limited perception of what governmental behavior is all about (Richard G. Lipsey and Kelvin Lancaster). The public sector involves, legitimately, both threat and integrative relationships. The grants economics literature shows that taxes, transfers, and expenditures take on an expanded meaning when viewed as something other than voluntary exchanges. Nonmarket failure takes its place as unsatisfactory behavior in threat and/or integrative relationships.

Defining this unsatisfactory behavior, at the root, is a problem of the relationship between the positive and normative elements of economics. This topic is important to environmental policy because the adoption of the polluter pays principle involves conflict rather than harmony. In this matter there are no value-free policy proposals nor are there enduring "solutions." Similar conditions obtain for the bulk of public policies involving taxes, transfers, and expenditures. Grants economics is openly normative and it needs an openly normative theory of nonmarket failure. Just as some public finance economists insist upon exchange model benefit taxation, there are environmental economists who have confused themselves and sometimes convinced others by making policy proposals, or stating principles for public policy, based on exchange economics. If such work is to be continued as something besides a purposefully sterile exercise, then the assumptions of the market model need to be made explicit and be proven to exist in reality. Given the current state of our knowledge about market theory, especially the theory of the second-best, it is fair to interpret market mechanism policy analysis as being either naive or purposefully deceptive. That is, the competitive model implies

harmony of interests and thus—when its assumptions are not met—constitutes a value premise. Values then become implicitly imbedded in both the diagnosis and the recommendation. From the perspective of grants economics, the problem is not one of determining whether or not a normative economics rabbit can be pulled from a positive economics hat, but clearly specifying the necessary value premises and supporting those chosen by reference to either ethical standards or prior public policy decisions. This is not to say that such reference is where the matter will end, quite the contrary. The arena of public policy has only been *entered*. It is the nature of these conflict situations that they do not have solutions. They do not have solutions because they are not closed-end puzzles. In the event that a legislative action on a policy matter fails to mollify one party at interest (or a coalition of parties), the contestants may turn to the judiciary. Now where normative policy is involved not only are economic *qua* exchange answers absent, but noneconomic solutions are lacking also. Rather than solutions or answers, the results of a legislative or judicial process are a set of working rules for an explicit societal agreement that is subject to change by reform or revolt. These arrangements are temporary. Though they may have quite long lives, they are not immortal. As values change, our perspective on public policy may be altered and, for example, effluent charges may become a more or less acceptable policy.

This mutability of values is the prime reason for the statement above which asked only consensus, rather than unanimity, in defining nonmarket failure. Nonmarket activity can only be judged by comparing public policies with moral standards and general societal working rules which have been previously established (John R. Commons). This approach to defining policy failure is compatible with the need for flexibility in policymaking. A policy may be unsatisfactory and be the wave of the future, the way of the past, or simply an aberration.

Using the language of grants economics, nonmarket failure is the breakdown of either the threat-submission system or the integrative system. Breakdown is judged by comparison with existing, though evolutionary, ethical standards and/or public policies.

## II. The Case of Effluent Charges

By examining many instances of failure, a set of common factors can be identified and a classification system developed. The reasons for nonmarket failure which emerge from an examination of policy critiques are these: (1) failure of structure, (2) failure of legitimacy, and (3) failure of information. Failure of structure is exemplified by policies which do not recognize the structure of federal, state, and local government and do not take into account the differential assignment of functional responsibility. Failure of legitimacy occurs when benefits of government activity accrue to only a few individuals rather than the public. In other words, where illegitimate redistribution occurs in the system. Failure of information brings about mistaken policy when: (a) the flow of information does not encourage foresight, (b) the flow of information is inappropriate to the organization's agenda, and (c) policies are pursued which have conflicting goals. In the space which remains each of these types of nonmarket failure will be discussed in terms of establishing and administering a set of effluent charges.

### *A. Failure by Structure*

Successful environmental policy must begin from an awareness of the structure of government. Local and/or state govern-

ment policies could be frustrated by federal initiatives. Effluent charges are too often prescribed without any clear notion of which level of government should levy such charges and whether or not such levies might have some relationship to other fundamental responsibilities of the levying unit, adjacent units, noncoterminous overlying units, or other units in the hierarchy. The merits of decentralization must be weighed not only against the ability to internalize externalities, but some reckoning with the other tasks of the unit can only be postponed and cannot be avoided. The levying of effluent charges could degenerate into a competitive situation among local governments or even states. Experience with competitive underassessment in the property tax field is instructive here. Assigning the responsibility for levying effluent charges to the unit with the real property assessment function offers the creative administrator some interesting opportunities for competing with other locales for manufacturing or service industry.

### *B. Failure by Legitimacy*

The distribution question, pursued by Martin Pfaff, is raised here in terms of its impact on public sentiment toward government. What sort of redistributional impact will be perceived as legitimate by the citizenry? We must not only try to answer the empirical question as to incidence, but we must also investigate the legitimacy of the impact. It seems likely that the populace would prefer that the polluter pay. There is evidence that economic literacy has progressed to the point that public discussions of burden shifting are possible. The public perception of the burden and the linkage of burden with specific products is bound to be an important issue. The public will likely consider illegitimate any policy that ignores burden shifting.

There is another dimension to the question of legitimacy. The public is easily convinced that the polluter should pay. Daily confrontation with the externalities generated by the behavior of others has made "polluter pays" a principle which is probably easier to agree upon than to achieve. The arrival at a consensus position is most important in policymaking, if not in model building, and economists ought to take heart and pursue the empirical work on incidence.

### *C. Failure by Information*

The possibility for failure by information seems to be a very real obstacle in the path of those who seek to develop a set of effluent charges. It is most important that data well beyond that generated by monitoring be available to decision makers. Information as to the changing composition of the effluent and the experiences of other governments levying similar charges needs to be considered. The levying of these charges can be expected to have something akin to income and substitution effects on the producer. Such information will permit a certain amount of foresight as to likely changes and likely responses to those changes. The flow of information must be appropriate to the agenda of the unit of government levying the charges. That is, information must be available to decision-makers which pertains to the relationship between the various public services provided (highways, education, etc.) and the effluent producing industry on which the charges are being levied. Certain variables are within the purview of local decision makers. If effluent charges are to accomplish a specific goal, then information must be available about other programs (such as the level of economic activity) so they may judge the impact of charges in the context of their other responsibilities. A related, but separable, issue is the availability of

information which might indicate that a government was pursuing conflicting goals. The levying units must be aware of the impact of effluent charges on economic growth and development so that a conscious decision about the socially tolerable level of emissions can be arrived at. Without this ability, effluent charges might be levied at too low a level and the community would be attracting polluters through this policy while their economic development efforts were being subverted. Goal conflicts are a considerable problem in many realms of policy making and the development of a set of effluent charges is not substantially different.

## REFERENCES

**F. M. Bator,** "The Anatomy of Market Failure," *Quart. J. Econ.*, Aug. 1958, 351–79, 351.

**K. E. Boulding,** *Economics as a Science*, New York 1970, 9–10.

**J. R. Commons,** *Legal Foundations of Capitalism*, Madison 1968, 331 ff.

**R. G. Lipsey and K. Lancaster,** "The General Theory of Second Best," *Rev. Econ. Studies*, 1956–57, *63*, 11–32.

# DISCUSSION

JANOS HORVATH, Holcomb Research Institute, Butler University: Projecting the searchlight of grants economics on environmental problems, we obtain a clearer perception of three things. Firstly, we can trace some implicit income and wealth transfers. Secondly, we can observe the mechanism by which such grants may affect productive efficiency and distributive equity. Thirdly, from this understanding we can design policy tools to induce environmentally desirable technological changes.

The papers of this panel have made commendable starts toward exploring what good and bad those grants can do to environmental quality and how they should be harnessed. I will attempt to examine to what extent their propositions have contributed to staking out the role of exchange versus grant transactions in environmental models.

What we ought to strive for is a broad enough framework within which all the various environmental policy problems could be integrated. In the realm of generalities, the first question is to detail the production and consumption functions of the economics of the coming spaceship earth, wherein the goal will no longer be to maximize output but rather to minimize throughput. Having perceived the spaceship earth, i.e., the environment, as the significant limiting factor, our task is to conduct human activities without wrecking ecological equilibrium. The fundamental environmental problem is, therefore, to achieve a reduction in the per capita use of the environment as a waste receptor.

In paving the way toward a comprehensive methodology of environmental policy problems, I recommend as building blocks, some pertinent concepts of grants economics. Herein I outline, for the sake of illustration, three scenarios.

*Scenario One: Implicit Grant to the Pollutor.* The pollutor gains through the non-abatement of the nuisance he causes, and in this sense he is the beneficiary of an implicit grant. The grantor is the inconvenienced neighbor, or nature to the extent that it absorbs the pollutant. The propensity to pollute is positively correlated with the capturable implicit grants. Alternatively, in a situation where the pollutor is no longer allowed to capture the implicit grant stemming from the non-abatement of external diseconomies, then his propensity to pollute is likely to decline.

The imposition of effluent charges will accomplish just that. Two of the papers at hand propose actions along those lines. Lloyd Orr proves quite convincingly that the halting of an implicit grant to the pollutor sets in motion technological innovations conducive to environmental improvement. Brumm and Dick show that subsidies for water treatment facilities are suboptimal or often counterproductive. In a sense, both papers argue against certain pathological grants which harm the environment. An exception to this is the transitional period toward the pollutor-pays-principle, when interregional dislocations may still warrant an elaborate network of explicit and implicit grants, as considered by Martin Pfaff.[1]

*Scenario Two: Implicit Grant from Nature to the Mineral Exploiter.* Natural resource exploitation is a composite economic transaction wherein the true value of the material is more than the cost of production (extraction). The ore itself is accepted as a gift from nature, which does not ask for a *quid pro quo* price. The exploiter becomes the recipient of an "implicit grant from nature." The fact that this is shared by the miner, the material processor, and the consumer, does not negate the idea that there is a grant from nature which, in turn, means that natural resources are undervalued. Certainly, the pivotal question is how to determine the size of this grant from nature via minerals. For this purpose we may make a distinction between the market value versus the surrogate value, and in a parallel fashion between the market price and the surrogate price. The market price is, of course, known. The other, the surrogate price, will approximate to a level where such large

[1] [Editor's note: I regret that Pfaff's paper for this session is not being published because the revised version did not reach my office in time. —R.F.]

quantities of the mineral would be produced that the supply function could become very elastic within the relevant range.

My proposition is substantiated by some recent advances in the science of mineral resources. The old school of natural resources held that minerals will sooner or later run out, or at least, that prices will rise sharply in the future. But the notion is naive because the entire planet is composed of minerals. There is evidence that vast quantities could be mined as either prices go up or as extraction costs go down pursuant to technological improvements. With only negligible exceptions, the volume of mineable materials could increase significantly—not by a factor of five or ten, but by a factor of 100 or 1000. Consequently, the problem boils down to finding that particular price level at which the supply function of the mineral continues to remain highly elastic throughout all possible ranges of demand shift. This surrogate price level serves then as the baseline for identifying the grant from nature, which will be the difference between the surrogate price and the prevailing market price. This latter idea has its antecedent in Ricardo's differential rent.

Considering, for example, that aluminum sold at 25-cents-a-pound price on ingot, and assuming that its surrogate price would be at 55 cents, then the grant from nature amounts to 30 cents. Since the global aluminum consumption from current mining was around 16 million tons during 1972, the world during this year was the beneficiary of an implicit grant from nature in the sum of about $9.6 billion. The corresponding sum for the United States would approximate to $3.2 billion.

Once perceived, the grant from nature may be employed as a policy tool in the resolution of various environmental problems. The level of pollution would decline if society reallocated the grant from nature by channeling part of it away from the heavy pollutor to the lesser pollutor; for instance, from bottle manufacturing to bottle recycling, or from copper to aluminum, as the case may be. The user in the socially-less-desirable or more polluting branch of the economy would receive the message in higher prices while the less objectionable branch would be signalled via lower prices.

*Scenario Three: Implicit Grants for Developing Solar Energy Technology.* Some futurologists envision the world's energy prospects as follows. During the next 15 years petroleum and gas will be utilized as the main source with simultaneously increasing reliance on coal. After about 25 years nuclear and solar energy will take the lead. Eventually a move away from fossil and even nuclear fuels is bound to ensue. What will cause this increased reliance on solar energy is that it constitutes a grant from the sun. By not satisfying more of its energy needs through the utilization of solar power—but capturing only a fraction of one percent beamed upon the earth—mankind foregoes an opportunity. Such a prodigious behavior is made possible through the availability of long accumulated solar energy, stored in the fossil fuels. Nuclear resources could of course offer new prospects, but the recurrent opposition to nuclear power calls for the consideration of alternative energy systems, such as solar energy.

As long as mankind continues to burn fossil fuels and uses up nuclear resources, it depletes a "stock grant" instead of living off a "flow grant." The long-run energy strategy can be viewed as a search for the ways to achieve a reduction in the use of stock grants while shifting to reliance on flow grants. A precondition to such a change in the utilization of energy resources is the development of suitable technology. In order to guide the direction and pace of technological change, a variety of policy tools are employed which are encompassed by the nomenclature of grants economics. Other things being equal, this optimal policy path would be to minimize marginal grant outlays per marginal energy output. This process also promises contributions to two other objectives: pollution abatement and fossil fuel preservation.

From the new technology we may expect lower costs so that solar power could become competitive in a wide range of uses. The downward move of the average cost functions and the outward shift of the supply functions will cause the price fall. As the economies of scale

are reaped, solar energy could compete with other alternatives. This stage is perhaps as near as five years in the heating of certain buildings and perhaps several decades away in the production of electricity and liquid hydrogen. The transition can be expected to be slower or speedier depending on the less or more judicious implementation of grant policies.

In concluding I should underscore that although the qualification of grant elements and the assessment of their leverage are novel techniques in economic analysis, explicit and implicit grants are as old as human communities. The economic history of the United States abounds with illustrations, such as the governmental underwriting of semiconductor research, civilian nuclear power, telestar, aviation, shipbuilding, trucking, or the railroadization a hundred years ago. Today's priority programs—the protection of the environment in general and the provision of energy and minerals in particular—may appear more formidable than any of the predecessors. Yet the measuring tools and normative principles of grants economics hold out the prospect for orchestrating a gamut of transactions about which Martin Pfaff wrote: cash grants, transfers in kind, concessionary leasing and lending, tax relief, relaxed rules, etc. Through quantifying the grant elements within each transaction and after calculating the tradeoffs between alternative choices, policy decisions could emerge with less uncertainty.

AMERICAN ECONOMIC ASSOCIATION

# PROCEEDINGS

OF THE

# EIGHTY-EIGHTH

# ANNUAL

# MEETING

DALLAS, TEXAS

DECEMBER 28-30, 1975

# THE JOHN BATES CLARK AWARD

*Citation on the Occasion of the Presentation of the Medal to*

DANIEL MCFADDEN

*December 29, 1975*

Daniel McFadden has been one of the most complete economists of his generation. He works in several different styles. He has written theoretical essays, worked on the development of new econometric tools, and undertaken large scale empirical inquiries. His investigations have ranged across production relations, decision making under uncertainty, development planning, welfare economics, and most recently urban transportation. There has been a certain uniform spirit behind all of his work—that the task of economists is to explain real economic phenomena. His work in pure and econometric theory has been conditioned by his appreciation of empirical reality. We honor him both for his seriousness about subject matter, and for the care and rigor which characterize all of his work.

# MINUTES OF THE ANNUAL MEETING
# DALLAS, TEXAS
# DECEMBER 29, 1975

The Eighty-eighth Annual Meeting of the American Economic Association was called to order by President Robert Aaron Gordon at 4:05 p.m., December 29, 1975, in the Dallas Convention Center.

The minutes of the meeting of December 29, 1974, were approved as published in the *American Economic Review*, May 1975, pages 441-47.

The Secretary (Rendigs Fels) presented the report of the Committee on Elections and the certification of the new officers for 1976 as follows:

In accordance with the bylaws on election procedures, I hereby certify the results of the recent balloting and report the actions of the Nominating Committee, the Electoral College, and the Committee on Elections.

The Nominating Committee, consisting of Robert Eisner, Chairperson, Bernard E. Anderson, Robert E. Gallman, Richard W. Parks, Leonard A. Rapping, Barbara B. Reagan, and Christopher A. Sims, submitted the nominations listed below for Vice-Presidents and members of the Executive Committee. The Electoral College, consisting of the Nominating Committee and the Executive Committee meeting together selected the nominee for President-elect. No petitions were received nominating additional candidates.

*President-elect*

Lawrence R. Klein

| *Vice-Presidents* | *Executive Committee* |
|---|---|
| Barbara R. Bergmann | Robert W. Fogel |
| Harry G. Johnson | Bert G. Hickman |
| Douglass C. North | Edmund S. Phelps |
| Albert E. Rees | Alice M. Rivlin |

The Secretary prepared biographical sketches of the candidates and distributed ballots in late summer. The Committee on Elections, consisting of Shirley K. Newton, Chairperson, Tom Holbrook, and Rendigs Fels, *ex officio*, canvassed the ballots and filed the results with the Secretary. From the Report of the Committee on Elections, I have the following information:

| | |
|---|---|
| Number of envelopes without names for identification | 250 |
| Number of envelopes received too late | 54 |
| Number of defective ballots | 36 |
| Number of legal ballots | 5,885 |
| | 6,225 |

On the basis of the canvass of the votes, I certify that the following persons have been duly elected to the respective offices:

*President-elect* (for a term of one year)
Lawrence R. Klein

*Vice-Presidents* (for a term of one year)
Barbara R. Bergmann
Harry G. Johnson

*Members of the Executive Committee* (for a term of three years)
Edmund S. Phelps
Alice M. Rivlin

In accordance with the actions of the Executive Committee at its meeting on March 14, 1975, amendments to Article I, Section 2; Article I, Section 3, Sentence 3; and Article V of the bylaws were submitted to the members in a mail ballot in conjunction with the balloting for officers. The ballots were canvassed by the Committee on Elections. On the basis of the canvass, I certify that the amendments were approved. As amended, Article I, Section 2, now reads:

2. There shall be nine classes of members other than honorary: regular members with the academic rank of assistant professor or lower or with annual incomes of $12,000 or less irrespective of rank, paying an annual fee of $25.00; regular members with the academic rank of associate professor or with annual incomes above $12,000 but not more than $20,000 paying an annual fee of $30.00 unless qualifying for a lower fee; regular mem-

bers with the academic rank of full professor or with annual incomes above $20,000, paying an annual fee of $35.00 unless qualifying for a lower fee; family members (two or more living at the same address, second membership without subscription to the publications of the Association) paying an annual fee of $5.00; junior members (available to registered students for three years only) paying an annual fee of $12.50; subscribing members paying an annual fee twenty percent higher than otherwise; contributing members paying an annual fee one hundred percent higher than otherwise; and life members comprising those who qualified for life membership by making a single payment of the designated amount prior to January 1, 1976, and exempt from annual fees.

As amended, Article I, Section 3, Sentence 3, now reads:

> In addition, the Executive Committee may elect additional Distinguished Fellows, but not more than two in any one calendar year, from economists of high distinction in the United States and Canada.

As amended, the previous text of Article V is now numbered 1 and the following has been added:

> 2. A resolution adopted at the annual meeting in which less than five percent of the membership of the Association has voted thereon shall be submitted within ninety days to a vote by mail ballot if a majority of the Executive Committee determines that, because of the nature or consequences of the resolution, all members should have the opportunity to participate in the final decision. Upon such a determination by the Executive Committee, the resolution shall cease to be effective and shall not become effective unless and until approved by a majority of the mail ballots cast. Mail ballots shall be accompanied by (1) a brief statement by the sponsor or sponsors in support of the resolution, (2) a brief statement of the views of the Executive Committee and, if the Committee favors or takes no position on the resolution, (3) a brief statement by a person or persons whom the Committee designates to represent opposing views.

In accordance with the action of the Executive Committee at its meeting of March 8, 1974, an amendment to Article I, Section 4, was submitted to a mail ballot in the spring of 1974. The ballots were canvassed by the 1975 Committee on Elections. On the basis of the canvass, I certify that the amendment was approved. As amended, Article I, Section 4, now reads:

> 4. Every member is entitled to receive, as they appear, the following publications of the Association: The *American Economic Review*, the *Journal of Economic Literature* and the *Directory of Members*.

A report by the Secretary-Treasurer had been distributed in advance of the meeting. He reviewed the report and moved adoption of the following resolution:

> BE IT RESOLVED that 80 percent of the dues of regular members with the rank of assistant professor or lower or with annual incomes of $12,000 or less irrespective of rank be considered payment for a year's subscription to both the *American Economic Review* and the *Journal of Economic Literature* and the same dollar amount of the payment by subscribers (nonmembers) and other regular members be considered payment for subscriptions to the *Review* and the *Journal*, with the remainder a contribution toward the Association's exempt functions, such as annual meetings and various economic research projects and programs. In the event that the Executive Committee under the authority of Article I, Section 2, changes the dues schedule and income brackets, the figure of $12,000 in this resolution will be considered to have been adjusted accordingly.

The motion was seconded and PASSED.

The president announced that Fels would be succeeded by C. Elton Hinshaw as Secretary, effective January 1, 1976, but would continue as Treasurer and Editor of the *Papers and Proceedings*. He then read the resolution adopted by the Executive Committee on December 27, 1975, and incorporated in its minutes.

The Managing Editors of the *American*

*Economic Review* (George H. Borts) and the *Journal of Economic Literature* (Mark Perlman) discussed their written reports, which also had been distributed in advance of the meeting and were to be published in the *Proceedings*. The Director of *Job Openings for Economists* (Hinshaw) read his report, likewise to be published in the *Proceedings*.

The Secretary presented the following resolutions, which were adopted unanimously:

> This meeting would like to record a special vote of thanks to Ralph T. Green, Chairman, and the other members of the National Convention Committee for their hard work, dedication and efficient management of the 1975 annual meetings of the Allied Social Science Associations.
>
> This meeting commends Franco Modigliani and the members of the Program Committee for planning a program intellectually stimulating, appropriately varied, and high in quality.

At this point, President Gordon introduced the new President, Franco Modigliani, who took the chair, commended his predecessor, and moved the thanks of the Association to him, a motion which was adopted by acclamation.

The new President gave an oral report on behalf of the Committee on Political Discrimination. David M. Gordon distributed copies of the statement, previously presented to the Executive Committee, entitled "Taking Political Discrimination Seriously" and proposing a resolution that would broaden the functions of the Committee on Political Discrimination. The President said that the statement would be referred to the Committee on Political Discrimination and subsequently would be considered by the Executive Committee. D. M. Gordon elaborated on the statement and the reasons for the proposed resolution. He explained that the resolution was considered not in order because it had not been submitted thirty days in advance of the annual meeting, but the reason was that the Committee on Political Discrimination had not finished its deliberations in time. David Barkin also spoke in favor of the statement and the resolution.

Oscar Goodman moved that future program chairmen of the American Economic Association not schedule any sessions in the future that conflict with the business meeting. The motion was seconded and PASSED. D. M. Gordon advocated steps to ensure larger attendance at the business meetings.

It having been pointed out that attendance at the business meeting was low, R. A. Gordon suggested that the Executive Committee at its March meetings deal with this problem. William Vickrey suggested that posters at the convention call attention to nonroutine items of business scheduled to come up at the business meeting. He also suggested that the business meetings be scheduled after the presidential address. Eric Bovet suggested that a quorum be established to force the Association to take steps to insure an adequate turnout. He also suggested that a new class of membership for retired persons be established.

There being no further business, the meeting was adjourned at 5:15 p.m.

RENDIGS FELS, *Secretary*

# MINUTES OF THE EXECUTIVE COMMITTEE MEETINGS

**Minutes of the Meeting of the Executive Committee in Washington, D.C., March 14, 1975.**

The first meeting of the 1975 Executive Committee of the American Economic Association was called to order at 9:20 a.m. on March 14, 1975, in the Washington Hilton Hotel, Washington, D.C. Present as members of the Executive Committee were R. A. Gordon, presiding, Carolyn Shaw Bell, George H. Borts, Andrew F. Brimmer, Martin Bronfenbrenner, Rendigs Fels, Walter W. Heller, Franco Modigliani, Mark Perlman, Charles L. Schultze, Paul M. Sweezy, and Burton A. Weisbrod. Absent were Irma Adelman, Kenneth J. Arrow, and Robert L. Heilbroner. Present for part of the meeting as members of the Nominating Committee were Robert Eisner, Chairperson, Bernard E. Anderson, Robert E. Gallman, Richard W. Parks, Leonard A. Rapping, Barbara B. Reagan, and Christopher A. Sims. Present for part of the meeting as members of the Committee on Honors and Awards were Lloyd G. Reynolds, Chairperson, John S. Chipman, and James W. McKie. Present as Chairperson of the Committee on the Status of Minority Groups in the Economics Profession was Marcus Alexis. Also present was the Assistant Secretary-Treasurer, C. Elton Hinshaw.

*Minutes.* The minutes of the meeting of December 27, 1974, were approved as corrected.

*Report of the Secretary* (Fels). The Secretary reported that a resolution was needed governing the application of retirement benefits to Association employees. He was authorized to approve such a resolution and put it into effect. The Secretary reported that, in response to the directive of the Executive Committee adopted at its meeting of December 27, 1974, the dates of the 1975 meetings in Dallas had been shifted from October 3–5 to December 28–30. He pointed out that the Executive Committees of different years had made three different decisions on the dates of the 1975 meetings and had made six decisions involving four changes in the site and dates of the 1976 meetings. In response to his contention that no Secretary, present or future, could live with a decision-making process leading to such vacillation, the Executive Committee VOTED to adopt the following procedure for determining sites and dates of future meetings: after consulting with the Executive Committee, the Secretary will propose the site and date, subject to the approval of the Executive Committee; decisions once made shall not be changed except in extreme cases; if a proposal of the Secretary is disapproved, the Executive Committee will not make some different decision but will instruct the Secretary to come in with a new proposal. In subsequently rejecting the Secretary's recommendation that the 1982 meetings be held in New York City on September 9–11, it made clear that it wanted the meetings held in New York City on December 28–30, thereby adhering to the letter if not the spirit of the new procedures it had just adopted.

*Report of the Treasurer* (Fels). The Treasurer reported that the budget deficit for 1975 was now projected at $155,000, compared to $129,000 in 1974; that the cash deficit for 1975 was projected at $76,000 compared to $119,000 the pre-

vious year; and that the net worth of the Association had fallen from $162,000 at the end of 1973 to a negative figure ($10,000) at the end of 1974 but was expected to rise to $54,000 by the end of the current year. Not surprisingly, the Executive Committee found these figures puzzling as well as disturbing. Considerable dissatisfaction having been expressed with the form of the financial reports, the Budget Committee was asked to reconsider budget procedures, including the investment income formula, and to report to the Executive Committee next December. The Treasurer pointed out that the 1975 projections reported above included $3,000 not yet authorized for the expenses of the Committee on Political Discrimination, which had been appointed in response to a resolution passed at the annual meeting. It was VOTED to appropriate $3,000 for the Committee on Political Discrimination. It was VOTED to approve and submit to the membership for ratification changes in the bylaws effective January 1, 1976, that would introduce progressivity in the dues structure in the following way: (1) assistant professors and those with lower academic rank would pay $25 per year, as would all members with annual incomes of $12,000 per year or less irrespective of rank; (2) associate professors and those with incomes above $12,000 but not more than $20,000 would pay $30 per year unless they qualified for the $25 rate under (1) above; and (3) full professors and those with incomes above $20,000 would pay $35 a year unless qualifying for a lower rate (all rates subject to indexation). It was then VOTED to approve and submit to the members for ratification a change in the bylaws raising the dues of family members to $5 per year effective January 1, 1976. It was also VOTED to approve and submit to the members for ratification a change in the bylaws abolishing the category of life member effective January 1, 1976, except for those already holding such membership. Finally it was VOTED to establish a voluntary page charge of $25 for articles in the regular issues of the *American Economic Review* (excluding the *Papers and Proceedings* issue) and the *Journal of Economic Literature*. The Treasurer was asked to make a greater effort to solicit advertising.

*Report of the Editor of the American Economic Review* (Borts). The Editor announced that he expected to reduce costs $12,000 a year by limiting the number of text pages per issue to 300 and eliminating free reprints. On recommendation of the Editor, it was VOTED to institute an extra charge for foreign mailings of the *Review* and the *Journal* approximately equal to the extra postage, effective January 1, 1976. It was VOTED to raise the *Review's* submission fee for nonmembers to $30, an extra $15 above the fee charged to members being applicable to dues if the author applies for membership.

*Report of the Editor of the Journal of Economic Literature* (Perlman). The Editor announced that the trim size of the *Journal* was being changed with savings estimated by the printer of $16,000 a year. He agreed to delete the index of authors of abstracts from the *Journal* for a saving of $1,400 a year. He reported on a proposal received from Knowledge Availability Systems Center (*KASC*). It was the sense of the meeting that he and the Secretary-Treasurer enter into negotiations with *KASC* and bring a proposal to the next meeting of the Executive Committee.

*Report of the Director of Job Openings for Economists* (Hinshaw). The Director reported that the excess of expenditures over revenues from subscriptions for 1974 was $12,000 and in 1975 was expected to be at least $10,000. He ex-

pressed the opinion that charges for listing vacancies should be instituted at a later time after the practice of open listing had become well established. He was asked to bring a proposal for such charges to the December meeting of the Executive Committee.

*Committee on Honors and Awards* (Reynolds). The Electoral College consisting of the Executive Committee and the Committee on Honors and Awards meeting together VOTED unanimously to award the John Bates Clark Medal to Daniel McFadden. At the suggestion of the Committee on Honors and Awards, the Executive Committee VOTED to approve a change in the bylaws, to be submitted to the members for ratification by mail ballot, removing the limit of twelve on the number of Distinguished Fellows in addition to ex-Presidents, thus permitting the election of two additional Distinguished Fellows annually.

*Nominating Committee* (Eisner). The Electoral College consisting of the Nominating and Executive Committees meeting together chose Lawrence R. Klein as the nominee for President-elect, Moses Abramovitz as Distinguished Fellow, and Robert Eisner as representative to the Social Science Research Council. The Chairperson of the Nominating Committee reported the following nominees for other offices in the 1975 election: for Vice President (two to be chosen), Barbara R. Bergmann, Harry G. Johnson, Douglass North, Albert Rees; members of the Executive Committee (two to be chosen), Robert W. Fogel, Bert Hickman, Edmund S. Phelps, Alice M. Rivlin.

*Committee on the Status of Minority Groups in the Economics Profession* (Alexis). The Chairperson distributed a written summary report on the Summer Program for Minority Students. In separate motions, it was VOTED (1) to approve a second Summer Program for Minority Students to be held in 1975, (2) to authorize the expenditure of $5,000 for the expenses of the Committee in 1975, (3) to underwrite the expenses of the 1975 Summer Program by a maximum of $5,000, and (4) to authorize the Committee to raise money for a fellowship program for graduate study in economics for minority groups.

*National Economic Association* (Gordon). The President reported on his conversations with the National Economic Association (*NEA*). It was agreed that the President in the future would invite suggestions of that organization for appointments to the Committee on Minority Groups in the Economics Profession. It was also agreed that the distinguished economists on the panel for the Visiting Scholars Program would be asked if they would be willing to help young black economists with their research and to make the list of those responding positively available to *NEA*.

*Committee on the Status of Women in the Economics Profession* (Reagan). The Chairperson reported that the size of the Committee has been reduced to seven from ten (with at least one male member).

*Committee on Honorary Members* (Fritz Machlup). Modigliani presented recommendations he had received by telephone from the Chairperson, who was prevented from attending by illness. The following foreign economists were elected Honorary Members: Edmond Malinvaud, Friedrich A. von Hayek, Don Patinkin, Trygve Haavelmo, and Nicholas Kaldor. No action was taken on the Committee's proposal to raise the maximum number of honorary members from 25 to 35, because the Executive Committee wanted more information about the reasons for the particular number recommended. The Secretary was asked to inform the Committee on Honorary

Members of its interest in electing economists from the Third World.

*Report of the Chairperson of the Program Committee* (Modigliani). The Chairperson reported on his plans for the program of the 1975 meetings.

*Resolutions Procedures* (Gordon). The President pointed out that a very small percent of the membership can pass important resolutions at the annual meetings. It was VOTED to approve and submit to the members for ratification by mail ballot a change in the bylaws under which resolutions passed at a meeting at which less than 5 percent of the members vote would be sent to all members for a vote by mail. It was agreed that the President-elect would consult the Counsel about what resolutions should be exempt from the requirement of a mail vote and about the wording of the proposed bylaw, which will be circulated to the Executive Committee prior to being submitted to the members.

*Expiration of Term of Secretary-Treasurer.* In view of the fact that the term of the Secretary-Treasurer expires at the end of 1975, the President announced that he would appoint a committee with broad terms of reference to bring a recommendation to the December meeting of the Executive Committee.

*Date of Spring Meeting.* It was agreed that the Executive Committee would meet on March 19–20, 1976, in Washington, D.C.

The meeting adjourned at 9:20 p.m.

**Minutes of the Meeting of the Executive Committee in Dallas, Texas, December 27, 1975.**

The second meeting of the 1975 Executive Committee of the American Economic Association was called to order in the Statler Hilton Hotel, Dallas, Texas, at 10:15 a.m. on December 27, 1975. The following members were present: R. A. Gordon, presiding, George H. Borts, Andrew F. Brimmer, Martin Bronfenbrenner, Rendigs Fels, Walter W. Heller, Franco Modigliani, Mark Perlman, Charles L. Schultze, and Burton A. Weisbrod. Also present were newly elected members Barbara R. Bergmann, Harry G. Johnson, Lawrence R. Klein, and Edmund S. Phelps. Absent were Irma G. Adelman, Kenneth J. Arrow, Carolyn Shaw Bell, Robert L. Heilbroner, Paul M. Sweezy, and newly elected member Alice M. Rivlin. Present as guests for parts of the meeting were Marcus Alexis, David M. Gordon, Irvin L. Grimes, C. Elton Hinshaw, Fritz Machlup, Barbara MacPhee, Barbara B. Reagan, and Lloyd G. Reynolds.

*Minutes.* The minutes of the meeting of March 14, 1975, were approved.

*Report of the Secretary* (Fels). The Secretary having expressed dissatisfaction with procedures for selecting sites and dates for the annual meetings, it was VOTED that:

> The Executive Committee should set general guidelines the Secretary should consider in proposing dates and places of annual meetings. Working within these guidelines the Secretary should make specific recommendations regarding future meeting times. (It is assumed that normally firm commitments will not extend beyond five years. Once approved by the Executive Committee, the place and dates of a particular annual meeting cannot be changed except under unusual circumstances, and a positive vote of two-thirds of the voting members of the Executive Committee would be required to effect a change.)

On recommendation of the Secretary, the Executive Committee VOTED to continue affiliation of the American Economic Association with the American Association for the Advancement of Science.

*Report of the Treasurer* (Fels). It was VOTED to approve the budget submitted

by the Treasurer and the Budget Committee subject to moderate changes. (For details of the budget, see the Treasurer's Report in the *Proceedings.*) It was agreed that a spcial committee would be appointed to reconsider the investment income formula.

*Report of the Editor of the American Economic Review* (Borts). On recommendation of the Managing Editor, the following persons were elected to the Board of Editors of the *American Economic Review* for three-year terms: Irma G. Adelman, David Baron, Robert Barro, Laurits R. Christensen, David Laidler, and Frank Stafford. The Editor was asked to reconsider the method of assessing page charges and report to the next meeting of the Executive Committee. It was suggested that the Editor commission an objective review article on the Means-Stigler controversy.

*Report of the Editor of the Journal of Economic Literature* (Perlman). On recommendation of the Managing Editor, the following persons were elected to the Board of Editors of the *Journal of Economic Literature:* George Feiwel, James Kindahl, Michio Morishima, and Charles Z. Wilson. It was VOTED to approve on a one-year trial basis an agreement with the Knowledge Availability Center of the University of Pittsburgh whereby the Association will receive 2 cents for each bibliographical reference based on the *Journal's* files and sold by the Center.

*Report of the Director of Job Openings for Economists* (Hinshaw). The Director reported that subscription revenues were very nearly equal to full costs of *Job Openings for Economists.* He therefore recommended no changes in charges. It was agreed that, in view of the early dates of the 1976 annual meetings, the possibility of a placement meeting late in 1976 or early in 1977 be explored.

*1976 Program Chairman* (Klein). The Chairman gave a brief report on plans for the 1976 program.

*U.S.-Soviet Exchanges* (Reynolds). It was VOTED to authorize the Committee on *U.S.*-Soviet Exchanges to solicit travel funds for American economists to attend the colloquium planned for June 1976 in Moscow. It was agreed that planning for a visit for Soviet economists to the *United States* in 1977 be postponed until after the 1976 visit by Americans to the *USSR.*

*Political Discrimination.* In the absence of Kenneth J. Arrow, Chairman of the Committee on Political Discrimination, Modigliani telephoned him and obtained an oral report. David M. Gordon distributed a statement entitled "Taking Political Discrimination Seriously" proposing a resolution asking that the scope of the Committee be broadened. D. M. Gordon requested that the resolution be considered either by the Executive Committee or by the membership at the annual meeting. R. A. Gordon explained that, since the Committee's functions had been determined at the 1974 annual meeting, the Executive Committee could not change its charge, and, since the resolution had not been submitted thirty days in advance, it would be out of order at the annual meeting of December 29, 1975. After suggesting that the resolution be referred to the Committee on Political Discrimination, he turned the chair over to Modigliani. Bronfenbrenner moved that an exception be made in this instance to the thirty-day rule. The motion died for lack of a second. After further discussion, it was VOTED that Modigliani would make an oral report for the Committee on Political Discrimination at the annual meeting, after which the floor would be open for discussion of the subject. It was agreed that the resolution would also be referred to the Committee on Political Discrimination. R. A. Gordon resumed the chair.

*Minority Group Economists* (Alexis).

The Executive Committee VOTED to carry forward $10,000 previously appropriated for the Committee on the Status of Minority Groups in the Economics Profession less amounts already spent and to appropriate an additional S5,000 with the hope and expectation that the funds would not be used. It was VOTED to pay the travel expenses of Alexis to Dallas.

*Women* (Reagan). It was VOTED to appropriate $9,700 for the Committee on the Status of Women in the Economics Profession for 1976. The Committee was asked to consider soliciting voluntary contributions from those associated with its activities with a view toward ultimate financial independence.

*Secretary-Treasurership* (Heller). On behalf of the Ad Hoc Committee on the Secretary-Treasurership (William H. Nicholls, George J. Stigler, and Walter W. Heller, Chairman), Heller moved that for the three-year term beginning January 1, 1976, C. Elton Hinshaw serve as Secretary and Rendigs Fels as Treasurer and Editor of the *Papers and Proceedings* (the division of duties between them to be as specified in the letter of October 10, 1975, from Fels to Heller). The motion was seconded and PASSED. It was then VOTED to approve the following resolution:

> First, that on the occasion of Rendigs Fels' decision not to be a candidate for reappointment as Secretary of the Association, the Executive Committee on behalf of the *AEA* membership expresses its warmest appreciation for the outstanding contribution he has made to the Association through his perceptive, judicious, and efficient management of the Association's affairs during an unusually demanding and difficult period.
>
> Second, that we consider ourselves most fortunate that Rendigs Fels is willing to continue to serve as Treasurer and Editor of the *Proceedings* for the coming three-year term.

*Publications*. In the absence of the Chairman of the Committee on Publications, Modigliani and Machlup reported on plans for a questionnaire surveying readers' opinions of the journals.

*International Economic Association* (Machlup). The question was raised of whether the American Economic Association should withdraw from the International Economic Association (*IEA*) in view of the latter's connection with *UNESCO* and *UNESCO*'s treatment of Israel. Machlup expressed the opinion that such action would hurt the Israelis, who regularly participate in *IEA* conferences, pointing out that Arabs seldom take part. Although this view did not command unanimous support, no action was taken.

*Bylaws* (Fels). It was VOTED to submit to the members for approval by mail ballot the following amendments to the bylaws:

*Change Article I, Section 2, to read:*

> 2. There shall be six classes of members other than honorary: regular members with the academic rank of assistant professor or lower or with annual incomes of $12,000.00 or less irrespective of rank, paying the base fee defined below; regular members with the academic rank of associate professor or with annual incomes above $12,000.00 but not more than $20,000.00 paying one and one-fifth times the base fee; regular members with academic rank of full professor or with annual incomes above $20,000.00 paying one and two-fifths times the base fee; family members (two or more persons living at the same address, second membership without subscription to the publications of the Association) paying one-fifth of the base fee; junior members (available to registered students for three years only) paying one-half the base fee; and life members comprising those who qualified for life membership by making a single payment of the designated amount prior to January 1, 1976, and exempt from annual fees.
>
> Effective January 1, 1976, the base fee is $25.00 per year. The Executive Com-

mittee may increase the dues schedule, including both the base fee and the income brackets for regular members, in proportion to the increase occurring after January 1, 1976 in relevant price and wage indexes, provided that the increase in any year shall not exceed ten percent.

*Change Article I, Section 3, to read:*

3. Foreign economists of distinction, not exceeding twenty-five in number, may be elected honorary members of the Association. Past presidents of the Association and members who have been awarded the Walker Medal shall be Distinguished Fellows. In addition, the Executive Committee may elect additional Distinguished Fellows, but not more than two in any one calendar year, from economists of high distinction in the United States and Canada. Candidates for Distinguished Fellowships shall be nominated by the Nominating Committee or the Executive Committee, and they shall be elected by the combined vote of the two committees. The Nominating Committee shall solicit and give due consideration to the recommendations of the Committee on Honors and Awards. The Nominating Committee is free to make no nominations in any particular year. However, it is not limited as to the number of candidates it may nominate in any year. Election to Distinguished Fellowship does not preclude election to any office of the Association.

*Change Article II to read:*

The Board of Trustees shall be composed of the voting members of the Executive Committee.

*Delete Article III, Section 2.*

*Renumber Article III, Section 3, and change it to read:*

2. The Association shall have the following officers who shall be appointed by ing officers who shall be appointed by the Executive Committee: a Secretary, a Treasurer, and a Managing Editor of the *American Economic Review,* a Managing Editor of the *Journal of Economic Literature,* and a Counsel. The terms of office of each of these officers shall be three calendar years.

*Renumber Article III, Section 4, to become Section 3.*

*Change Article IV, Section 2, to read:*

2. Before October 1 of each year, the President-elect of the Association shall appoint a Nominating Committee for the following year, this Committee to consist of a past officer as Chairman and not less than five other members of the Association. In addition to appointees chosen by the President-elect, the Committee shall include any other member of the Association nominated by petition including signatures and addresses of not less than 2 percent of the members of the Association, delivered to the Secretary before December 1. No member of the Association may validly petition for more than one nominee for the Committee.

The names of the Committee shall be announced to the membership immediately following its appointment and the membership invited to suggest nominees for the various offices to the Committee. The Nominating Committee for each year shall be instructed to present to the Executive Committee on or before March 31 a nominee for the President-elect and two or more nominations for each other elective office to be filled, except the presidency, all these nominees being members of the Association. The members of the Nominating and Executive Committees shall constitute an Electoral College which shall consider the nominee of the Nominating Committee for the President-elect and select a single candidate for that office. In the voting in the Electoral College, each member shall have one vote provided that the number of members of the Nominating Committee present does not exceed the number of members of the Executive Committee present; otherwise the members of the Nominating Committee present shall have fractional votes such that their sum equals the number of members of the Executive Committee present.

The Secretary shall inform all members of the Association of the actions of the Nominating Committee and the Electoral College not later than the June issue of the *American Economic Review.* An additional nomination for any office may be

made by petition, delivered to the Secretary by August 1, including signatures and addresses of not less than 6 percent of the membership of the Association for the office of President-elect and not less than 4 percent for each of the other offices. No member of the Association may validly petition for more than one nominee for the Executive Committee, one nominee for Vice-President, and one nominee for President-elect.

The election of officers by the membership shall take place by a mail ballot conducted by the Secretary each year. The ballot shall list all nominees alphabetically, with indication "nominated by petition" where applicable. Space shall be provided on the ballot for the individual voter's alternative choices for all offices. The Secretary shall mail the ballots to all members as soon as practicable after August 1 and set a deadline for receipt of ballots in the Secretary's office no earlier than October 1 and no later than November 30. The candidates with the highest number of votes for the various offices will be elected. The results of the election shall be certified and announced by the Secretary at the annual business meeting or in the *American Economic Review*.

*Renumber Article IV, Section 3, to become Section 4, and add as Section 3:*

3. The President-elect shall be responsible for the program for the annual meeting of the year in which he serves. He may at his discretion appoint a Program Committee to assist him.

*Renumber Article IV, Section 4 to become Section 5.*

*Renumber Article IV, Section 5, and change to read:*

6. The Executive Committee shall have the control and management of the funds of the corporation. It may fill vacancies in the list of officers, and may adopt any rules or regulations for the conduct of its business not inconsistent with this constitution or with rules adopted at the annual meetings. It shall act as a committee on time and place of meetings and perform such other duties as the Association shall delegate to it. A quorum shall consist of five voting members.

*Renumber Article IV, Section 6, and change to read:*

7. The Managing Editors shall, with the advice and consent of the Executive Committee, appoint members of Editorial Boards to assist them. They shall be *ex officiis* members and chairpersons of their respective Boards, which shall have charge of the publication of the *American Economic Review* and the *Journal of Economic Literature*.

*Renumber Article IV, Section 7, to become Section 8.*

*Change Article V, Section 1, to read:*

1. The annual meeting of this corporation shall be held at such time and place as may be determined by the Executive Committee. Notice of such time and place shall be given by publication in the *American Economic Review,* at least three months prior to such meeting.

*Change Article V, Section 2, to read:*

2. A resolution adopted at an annual meeting in which less than five percent of the membership of the Association has voted thereon shall be submitted within ninety days to a vote by mail ballot if a majority of the Executive Committee determines that, because of the nature or consequences of the resolution, all members should have the opportunity to participate in the final decision. Upon such a determination by the Executive Committee, the resolution shall cease to be effective and shall not become effective unless and until approved by a majority of the mail ballots cast. Mail ballots shall be accompanied by (1) a brief statement by the sponsor or sponsors in support of the resolution, (2) a brief statement of the views of the Executive Committee and whether the Committee favors, opposes, or takes no position on the resolution, and (3) a brief statement by a person or persons whom the Committee designates to represent opposing views.

*Change Article VI to read:*

Amendments, after having been approved by a majority of the Executive Committee, may be adopted by a majority of votes cast in a mail ballot.

*Richard D. Irwin, Inc.* (Grimes). The Vice Chairman of Richard D. Irwin, Inc., proposed a settlement of the Association's claim for $34,000 under which the Association would receive $10,000 and all the remaining copies of the books in the foreign translation series (about 13,500 copies in all) to be shipped to any destination at the expense of the Association. Grimes was asked to investigate the possibilities of (1) bulk remaindering of the entire stock and (2) remaindering of the stock by the Association with Richard D. Irwin, Inc., holding it and filling orders at a cost estimated at 50 cents per copy. It was agreed to take the matter up again at the next meeting of the Executive Committee.

*Expiration of Terms of the Managing Editors.* It was VOTED that the incoming President appoint a small committee to make a report in March 1976 on reappointment of the present Managing Editors, whose terms expire at the end of 1977, or appointment of new ones.

The continuing members of the Executive Committee having expressed thanks to the retiring members for their service, the meeting was adjourned at 10:10 p.m.

RENDIGS FELS, *Secretary*

# REPORT OF THE SECRETARY FOR 1975

*Annual Meetings.* In 1976 the annual meetings will be held at the Chalfonte-Haddon Hall in Atlantic City on September 16–18, with the employment service beginning operations on the 15th. This is a departure from the traditional Christmas holiday dates. The departure resulted from a referendum of the membership in 1970, which could not be acted on earlier because hotel accommodations for large meetings need to be reserved five to eight years in advance. (In the case of the 1975 meetings, in Dallas, it had proved possible to change the dates from Christmas to early October, but the decision to do so was reversed by the Executive Meeting at its meeting on December 27, 1974.)

The schedule for meetings after 1976 is: December 28–30, 1977, in New York City with headquarters at the New York Hilton Hotel; August 29–31, 1978, in Chicago with headquarters at the Conrad Hilton Hotel; December 28-30, 1979, in Atlanta with headquarters at the Atlanta Hilton; September 5–7, 1980, in Denver with headquarters at the Denver Hilton Hotel; and December 28–30, 1981, in Washington, D.C. with headquarters at the Washington Hilton Hotel and the Sheraton-Park Hotel. No decision has been made for 1982. The Executive Committee has made a tentative decision to meet in San Francisco December 28–30, 1983.

*Membership.* The total number of members and subscribers, shown in Table 1, reached an all-time high of 26,787 at the end of 1975. The previous peak had been at the end of 1970, just before dues and subscription rates were doubled. By the end of 1972, the number had fallen from 26,173 to 24,489. Since then the total has been rising. The number of subscribers is still slightly below 1970, reflecting the greater increase in subscription rates than dues in the 1970–75 period. The number of regular (annual) members is also a little lower, the increase in the total having come in the minor categories of membership.

TABLE 1—MEMBERS AND SUBSCRIBERS
(End of year)

| | 1970 | 1972 | 1975 |
|---|---|---|---|
| Class of Membership | | | |
| Annual | 16,075 | 14,462 | 16,011 |
| Junior | 2,077 | 1,886 | 2,367 |
| Family | 289 | 262 | 335 |
| Complimentary | 160 | 345 | 449 |
| Life | 291 | 315 | 383 |
| Honorary | 16 | 72 | 19 |
| Total Members | 18,908 | 17,342 | 19,564 |
| Subscribers | 7,265 | 7,147 | 7,223 |
| Total Members and Subscribers | 26,173 | 24,489 | 26,787 |

*Permission to Reprint and Translate.* Official permissions to quote from, reprint, or translate and reprint articles from the *American Economic Review* and the *Journal of Economic Literature* totaled 204 in 1975 compared to 180 in 1974. In addition, permission was granted to reprint 263 abstracts from the *Journal.* Upon receipt of a request for permission to reprint an article, the publisher or editor making the request is instructed to get the author's permission in writing and send a copy to the Secretary as a condition for official permission. The Association suggests that authors charge a fee of $150, but they may charge some other amount, enter into a royalty arrangement, waive the fee, or refuse permission altogether.

*Visiting Scholars Program.* The Visiting Economics Scholars Program, modeled on the Visiting Scientists Program in Economics formerly supported by the National Science Foundation, was inaugurated in the academic year 1974–75 and has continued under the direction of C. Elton Hinshaw. Its purpose is to facilitate visits by leading economists to smaller colleges emphasizing teaching. During 1974–75 there were eight such visits.

*Committees and Representatives.* Listed below are those who served the Association during 1975 as members of committees or as representatives. Years in parentheses indicate the final year of the term to which they most recently have been appointed. On behalf of the Association, I wish to thank them all for their service.

### *Standing Committees*

ADVISORY COMMITTEE ON THE HISTORY OF THE ASSOCIATION

George J. Stigler, *Chairperson*
Joseph Dorfman
Harold F. Williamson, *Corresponding Secretary*

ADVISORY COMMITTEE ON STUDIES OF THE LABOR MARKET FOR ECONOMISTS

F. Ray Marshall, *Chairperson*
Barbara Reagan
T. Aldrich Finegan
Francis M. Boddy

BUDGET COMMITTEE

Irma Adelman, *Chairperson* (1975)
Andrew F. Brimmer (1976)
Burton A. Weisbrod (1977)
Rendigs Fels, *Ex Officio*
R. A. Gordon, *Ex Officio* (1975)
Franco Modigliani, *Ex Officio* (1976)

CENSUS ADVISORY COMMITTEE

Richard Ruggles (1978), *Chairperson* (1975)
Joseph Burns (1975)
Anne P. Carter (1978)
Dean E. McKee (1975)
J. Fred Weston (1975)
Gardner Ackley (1976)
Armen Alchian (1976)
Andrew F. Brimmer (1976)
Anthony Downs (1976)
James R. Nelson (1976)
Jacob Mincer (1977)
George L. Perry (1977)
Lee Preston (1977)
Phyllis Wallace (1977)
Dale Jorgenson (1977)
*Barbara Bergmann (1978)
*Robert Lanzillotti (1978)
*William Niskanen (1978)

COMMITTEE ON ECONOMIC EDUCATION

G. L. Bach, *Chairperson* (1976)
Elisabeth Allison (1978)
Kenneth E. Boulding (1975)
Allen C. Kelley (1975)
Henry H. Villard (1976)
Phillip Saunders (1977)
Rendigs Fels, *Ex Officio*

COMMITTEE ON HONORARY MEMBERS

Fritz Machlup, *Chairperson* (1976)
Lawrence R. Klein (1976)
Bent Hansen (1978)
W. Arthur Lewis (1978)
Leonid Hurwicz (1980)
Paul A. Samuelson (1980)

* Newly appointed in 1976

COMMITTEE ON HONORS AND AWARDS
Lloyd G. Reynolds, *Chairperson* (1976)
Gary Becker (1976)
Irma Adelman (1978)
Marcus Alexis (1978)
John Chipman (1980)
James W. McKie (1980)

COMMITTEE ON THE STATUS OF MINORITY GROUPS IN THE ECONOMICS PROFESSION
Marcus Alexis, *Chairperson* (1977)
Bernard E. Anderson (1975)
Karl Gregory (1975)
Phyllis Wallace (1975)
George Borts (1976)
Andrew Brimmer (1976)
Alice Rivlin (1976)
James Tobin (1977)
Charles Z. Wilson (1977)

COMMITTEE ON POLITICAL DISCRIMINATION
Kenneth J. Arrow, *Chairperson*
William J. Baumol
John G. Gurley
Anne Krueger
F. Ray Marshall
Carl Stevens
Thomas E. Weisskopf

COMMITTEE ON PUBLICATIONS
William Brainard (1977), *Chairperson* (1975)
John G. Gurley (1978)
Robert Heilbroner (1975)
Peter Diamond (1976)
Robert Lampman (1976)
Michael Lovell (1977), *Chairperson* (1976)
*Robert Ferber (1978)
*Robert Gallman (1978)
*C. Elton Hinshaw, *Ex Officio*

* Newly appointed in 1975

COMMITTEE ON THE STATUS OF WOMEN IN THE ECONOMICS PROFESSION
Barbara Reagan, *Chairperson* (1976)
Francine Blau (1975)
Florence Weiss (1975)
Walter W. Heller (1976)
Isabelle Sawhill (1977)
Janice Madden (1977)
Nancy H. Teeters (1977)

FINANCE COMMITTEE
Beryl W. Sprinkel, *Chairperson*
C. Wells Farnham
Rendigs Fels
Milton Friedman
Walter W. Heller

ECONOMICS INSTITUTE POLICY AND ADVISORY BOARD
Edwin S. Mills, *Chairperson* (1978)
Irma G. Adelman (1975)
Walter P. Falcon (1976)
Daniel Schydlowsky (1976)
Arnold Harberger (1977)
Richard H. Holton (1977)
Paul G. Clark (1978)

JOINT COMMITTEE WITH THE ASSOCIATION OF AMERICAN LAW SCHOOLS
George J. Stigler, *Co-Chairperson* (1976)
Gerald M. Meier (1975)
Peter O. Steiner (1975)
Alvin Klevorick (1976)

U.S.-SOVIET EXCHANGES
Fritz Machlup, *Chairperson* (1974–75)
Lloyd G. Reynolds, *Chairperson* (1975–76)
Abram Bergson
John Meyer

## *Special Committees*

COMMITTEE ON COMPUTERIZATION
John R. Meyer, *Chairperson*

NOMINATING COMMITTEE (1975)
Robert Eisner, *Chairperson*
Bernard E. Anderson
Robert E. Gallman
Richard W. Parks
Leonard A. Rapping
Barbara Reagan
Christopher A. Sims

NOMINATING COMMITTEE (1976)
Kenneth Arrow, *Chairperson*
William James Adams
Duran Bell
Marianne Ferber
Jack Hirschleifer
Leonard Rapping
Vincent Tarascio

COMMITTEE ON ELECTIONS (1975)
Shirley K. Newton, *Chairperson*
Tom Holbrook
Rendigs Fels, *Ex Officio*

DALLAS LOCAL ARRANGEMENTS COMMITTEE (1975)
Ralph T. Green, *Chairperson*
Hilary S. Ball
James A. Byrd
Douglas Foster
Josef Hadar
S. T. Keim, Jr.
George W. Kelly
Barbara A. MacPhee
T. E. McMillan, Jr.
J. Carter Murphy
Thomas P. Robertson
Robert Smith, III
Jack Strickland
Paul Weathers
Eugene C. Zorn, Jr.

ATLANTIC CITY LOCAL ARRANGEMENTS COMMITTEE (1976)
Edward G. Boehne, *Chairperson*
Richard B. Benedict
Charles Campbell
John J. Clark
Roger D. Collons
Thomas J. Doty
Marge Epps
Nancy Foltz
Robert C. Forrey
Vince Franco
Alan Gart
Alan Gersh
Shirley Goetz
Ben Han
A. Gilbert Heebner
Kathleen C. Holmes
Gerrold T. Jacobs
Barbara MacPhee
Fred W. Malkin
Mary Murray
Andrew T. Schlenker
Joe Tumbler
Henry Watson
Bertham Zumeta

## *Council and Other Representatives*

AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE
Stephen Goldfeld (1976)

AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SLAVIC STUDIES
Janet Chapman (1976)

AMERICAN COUNCIL OF LEARNED SOCIETIES
William Parker (1978)

AMERICAN POLITICAL SCIENCE ASSOCIATION—JOINT RESEARCH PROJECT ON CONFIDENTIALITY OF RESEARCH SOURCES
Gary Fromm

FEDERAL STATISTICS USERS' CONFERENCE
John W. Kendrick (1977)

INTERNATIONAL ECONOMIC ASSOCIATION
Fritz Machlup (1975)
Abram Bergson (1978)

INTERSOCIETY COMMITTEE ON TRANSPORTATION
Charles A. Taff

JOURNAL OF RESEARCH ON CONSUMER BEHAVIOR
Kelvin J. Lancaster (1978)

NATIONAL ARCHIVES ADVISORY COUNCIL—GENERAL SERVICES ADMINISTRATION
Robert Gallman (1978)

NATIONAL BUREAU OF ECONOMIC RESEARCH
Willard L. Thorp (1975)
*Carl F. Christ (1978)

NATIONAL COUNCIL FOR ACCREDITATION OF TEACHER EDUCATION
Phillip Saunders (1977)

SOCIAL SCIENCE RESEARCH COUNCIL
William J. Baumol (1975)
Lawrence R. Klein (1976)
Guy Orcutt (1977)
Robert Eisner (1978)

* Newly appointed in 1976

## *Representatives of the Association on Various Occasions—1975*

INAUGURATIONS

Rev. Charles David Sherrer, Kings College
T. S. Saini

Jerald C. Walker, Baker University
Paul E. Junk

Samuel Alston Banks, Dickinson College
William F. Railing

William James Byron, University of Scranton
Robert E. Lucas, Wilmington College
Joseph T. Chao

Philip Harding Jordan, Jr., Kenyon College
Herbert S. Parnes

Manning Mason Pattillo, Jr., Oglethorpe University
Robert C. Vowels
Robert W. Rabald

# REPORT OF THE TREASURER FOR THE YEAR ENDING DECEMBER 31, 1975

Table 1 gives budget data for 1975 and 1976. The budgets are compared with the actual results for 1974 and with the results anticipated for 1975 at the time (December 27, 1975) that the 1976 budget was approved by the Executive Committee. The actual results for 1975 are given below in the financial statements following the Auditors' Report. Since the Treasurer's Report is sent to the printer before the audit takes place, the final results cannot be known at the time this report is written. Ordinarily actual results differ from anticipated by a significant margin. Nevertheless, it is clear that a substantial improvement in the financial position of the American Economic Association took place in 1975, thanks to actions taken by the Executive Committee at its meeting in March 1975, cost cutting by the managing editors of the two principal publications, and the rise in the stock market.

At the end of 1974, the net worth of the Association was negative by nearly $10,000. In March 1975, just before the meeting of the Executive Committee, the expected deficit for 1975 on the basis on which the Association's accounts have been kept during recent years was $155,000. (See Table 1.) Actions taken at the March meeting in response to this situation and recorded in the Minutes printed above, together with underspending by the managing editors, sharply reduced the 1975 deficit compared to what had been expected and brought the deficit expected for 1976 down to a very low figure. The deficits for 1975 and 1976 include provision for a reserve for a future *Directory* and part of the real capital losses incurred in 1973–74 but not recognized until 1975 and 1976. Without these items, there would be surpluses.

Table 1 is mainly on an accrual basis. Table 2 shows the cash budget for 1976. An increase in cash of $83,000 is expected. The first line in Table 2 repeats the deficit shown in Table 1. The next two items represent adjustments to eliminate the reserve for the *Directory* and the real capital losses which are included in Table 1 but do not involve a change of cash. The fourth item, the manufacturing cost of the four volumes of the *Index of Economic Articles* to be published in 1976, will use cash but will be recovered from sales in future years. Line 5 shows a substantial expected increase in deferred income associated with the increase in dues for high-income members which went into effect January 1, 1976. The Association collects dues and subscriptions in advance, which are shown on the books as deferred income, a liability. Whenever there is an increase in dues and subscription rates, the associated increase in deferred income generates cash.

The rise in the stock market in 1975 was faster than the rise in the *GNP* deflator and therefore generated real capital gains. The Association recognizes only one-third of the real capital gains or losses from equities in the year in which they occur, recognizing the other two-thirds in the income statements for the two subsequent years. Consequently, Table 1 does not fully reflect the gain in the value of the Association's portfolio in 1975 (nor do the figures for 1974 fully reflect the

TABLE 1—AMERICAN ECONOMIC ASSOCIATION BUDGET, ACCRUAL BASIS, FOR 1975

(Thousands of dollars)

| | 1976 Budget[a] | 1975 Budget[b] | 1975 Expected[c] | 1974 Actual |
|---|---|---|---|---|
| REVENUES | | | | |
| *Operating Income* | | | | |
| Dues and subscriptions | 662 | 620 | 602 | 569 |
| Advertising | 69 | 61 | 65 | 61 |
| Sales—Miscellaneous | 20 | 15 | 24 | 22 |
| Sales—Mailing list | 32 | 24 | 32 | 24 |
| Annual meeting | 15 | 10 | 15 | 17 |
| *Job Openings for Economists* (*JOE*) | 20 | 15 | 16 | 4 |
| Other income | 23 | 18 | 17 | 13 |
| Subtotal | 841 | 763 | 771 | 711 |
| *Investment Income* | | | | |
| Interest and dividends | 40 | 43 | 31 | 26 |
| Real capital gains (losses) | (60) | (99) | (86) | (82) |
| Subtotal | (20) | (56) | (55) | (56) |
| TOTAL REVENUE | 821 | 707 | 716 | 655 |
| EXPENSES | | | | |
| *Publications* | | | | |
| *American Economic Review* | 214 | 232 | 193 | 201 |
| *Journal of Economic Literature* and *Index* | 277 | 274 | 258 | 249 |
| *Papers and Proceedings* | 56 | 58 | 51 | 52 |
| *Directory*[d] | 50 | 50 | 55 | 69 |
| *JOE* | 15 | 15 | 13 | 16 |
| Subtotal | 612 | 629 | 570 | 587 |
| *Operating & Admin.* | | | | |
| Salaries | 120 | 123 | 112 | 99 |
| Rent | 8 | 8 | 8 | 8 |
| Committees | 25 | 25 | 20 | 13 |
| Other | 75 | 77 | 68 | 75 |
| Subtotal | 228 | 233 | 208 | 196 |
| TOTAL EXPENSES | 840 | 862 | 778 | 783 |
| SURPLUS (DEFICIT) | (19) | (155) | (62) | (129) |

[a] As approved by the Executive Committee, December 27, 1975. Does not include minor changes made subsequently in accordance with Executive Committee actions and policies.

[b] The projections of revenues shown in the 1975 budget are as revised in March in the light of the actual results for 1974. The estimates of expenditures take account of the actions by the Executive Committee at its meeting on December 27, 1974, and by the membership at the Annual Meeting of December 29, 1974. Actions taken by the Executive Committee at its meeting on March 14, 1975, however, are not included in the estimates of revenues and expenditures; they account for much of the disparity between the budget and the actual results.

[c] Estimates prepared in November 1975 of revenues and expenses for the entire year.

[d] Includes $50,000 in each of the years 1975 and 1976 set aside as a reserve for a future *Directory*. The $69,000 in 1974 and the other $5,000 in 1975 represent actual expenditures for the 1974 *Directory* in excess of the reserve accumulated previously.

TABLE 2—AMERICAN ECONOMIC ASSOCIATION CASH BUDGET FOR 1976
(Thousands of dollars)

| | |
|---|---|
| SOURCES (USES) OF CASH | |
| 1. Budget surplus (deficit) | (19) |
| 2. Addition to reserve for *Directory* | 50 |
| 3. Real capital losses (gains) | 60 |
| 4. (Manufacturing cost of *Index of Economic Articles*, 4 vols.) | (60) |
| 5. Increase (decrease) in deferred income | 39 |
| 6. Decrease (increase) in receivables | 2 |
| 7. Increase (decrease) in accounts payable and accrued liabilities | 12 |
| 8. (Furniture and equipment to be purchased) | (1) |
| 9. Depreciation | * |
| INCREASE (DECREASE) IN CASH | 83 |

* Less than 0.5.

losses). But the balance sheet (printed below following the Auditors' Report) shows the current value of the portfolio. The negative figure for net worth in the balance sheet at the end of 1974 resulted partly from the decline in the stock market during that year, and the rise in the stock market in 1975 is fully reflected in the net worth, which is now significantly positive. But it is low relative to annual expenditures and relative to the highest deficit ($242,000 in 1970) ever experienced.

The accumulated deficits for 1969–75 totaled nearly half a million dollars. With the net worth as low as it is, the Association must cease having deficits, and it needs to run surpluses for several years until the ratio of net worth to annual expenditures is built up to an adequate level. The change in the dues structure instituted at the beginning of 1976 is intended to achieve this result.

RENDIGS FELS, *Treasurer*

# REPORT OF THE FINANCE COMMITTEE*

The accompanying inventory summary lists the securities held by the American Economic Association as of December 31, 1975, with costs and market values as of that date. The total market value of the securities portfolio at year end was $490,672. After making adjustments for cash additions and withdrawals (including a sizeable withdrawal of $65,000 made from the portion of the Fund represented by the Ford Foundation grant), we estimate that the Association's investment portfolio (on a total return basis) increased in value by 30.6 percent during 1975. During the same period, the Standard & Poor's 500 Index provided a total return of +37.2 percent. Utilizing a longer term period of time (from the end of 1967 to December 31, 1975), which includes both recoveries and recessions and rising and falling securities markets, the account's total return was + 34.2 percent while the Standard & Poor's 500 Index experienced a total return of +23.4 percent and an unweighted average of all of the stocks on the New York Stock Exchange actually declined on a comparable basis during this period by 27 percent.**

It should be remembered that the $490,672 figure referred to above includes a Special Grant that was made by the Ford Foundation in January of 1969 and subsequently commingled with the Association's account. As of December 31, 1975, the Association's portion of the aggregate account was $390,022, or 79.5 percent and the Special Grant represented the remaining $100,650, or 20.5 percent of the total. These calculations are after taking the $65,000 withdrawal from the Special Grant into consideration.

During the past twelve months, several changes were made in the equity component of the account. The most important of the changes that were made involved the sale of 200 Atlantic Richfield, 2,200 Rank Organisation, 400 ERC, 1,000 Hoerner Waldorf, 800 Manufacturers Hanover, 4,601 Capital Opportunities Fund, and 400 Xerox; and new purchases of 400 Abbott Labs, 500 Philip Morris, and 600 Seven Up. It should be noted that a significant proportion of the sales listed above were undertaken because of the

* The Report of the Finance Committee is informational and is not an audited financial statement. Consequently, there may be some discrepancies between figures in the Report of the Finance Committee and the Auditors' Report which follows.

** The unweighted average of all the common stocks on the New York Stock Exchange was adjusted to include the yield on the Standard & Poor's 500 Index.

TABLE 1—INVENTORY SUMMARY AS OF DECEMBER 31, 1975

| | Value | Percent | Estimated Income |
|---|---|---|---|
| Cash Equivalents and Short-Term Securities | $ 56,941 | 11.6 | $ 3,425 |
| Medium-Term Securities | 0 | 0.0 | 0 |
| Long-Term Securities and Preferred Stocks | 0 | 0.0 | 0 |
| Convertible Securities | 0 | 0.0 | 0 |
| Equity Securities | 433,731 | 88.4 | 12,845 |
| Total | 490,672 | 100.0 | 16,270 |

TABLE 2—INVENTORY AND APPRAISAL AS OF DECEMBER 31, 1975

| | Amount | Price | Value | Unit Cost | Total Cost | Est. Income |
|---|---|---|---|---|---|---|
| CASH EQUIVALENTS AND SHORT-TERM SECURITIES (11.6 percent) | | | | | | |
| *Cash Equivalents (0–1 Year) (9.5 percent)* | | | | | | |
| Cash | | | $741 | | $741 | |
| Ford MTR NTS (5.875 03/04/76) | 36,000 | 100 | 36,000 | 100 | 36,000 | 2,115 |
| U.S. Treas Bills (04/08/76) | 10,000 | 100 | 10,000 | 98 | 9,785 | 510 |
| | 46,000 | | 46,000 | | 45,785 | 2,625 |
| Subtotal Cash Equivalents | | | 46,741 | | 46,526 | 2,625 |
| *Other Short-Term Securities (1–5 Years) (2.1 percent)* | | | | | | |
| U.S. Treasury (8.00 02/15/77) | 10,000 | 102 | 10,200 | 102 | 10,173* | 800 |
| Subtotal other Short-Term Securities | | | 10,200 | | 10,173 | 800 |
| Total Cash and Fixed Income Securities | | | 56,941 | | 56,699 | 3,425 |
| EQUITY SECURITIES (88.4 PERCENT) | | | | | | |
| *Utilities (4.0 percent)* | | | | | | |
| Central and Southwest | 1,000 | 17 | 17,250 | 8 | 8,316* | 1,160 |
| *Banks (12.1 percent)* | | | | | | |
| Citicorp | 800 | 30 | 23,600 | 29 | 23,440* | 704 |
| First Bank System | 700 | 41 | 28,700 | 20 | 13,831* | 1,064 |
| | | | 52,300 | | 37,271 | 1,768 |
| *Other Financial (4.3 percent)* | | | | | | |
| Alexander and Alexander | 600 | 31 | 18,600 | 19 | 11,175* | 540 |
| *Foods and Containers (10.6 percent)* | | | | | | |
| Philip Morris | 500 | 53 | 26,500 | 46 | 22,857* | 500 |
| Seven Up | 600 | 33 | 19,500 | 29 | 17,300* | 504 |
| | | | 46,000 | | 40,157 | 1,004 |
| *Mining and Metals (11.3 percent)* | | | | | | |
| MAPCO | 700 | 37 | 25,550 | 18 | 12,397 | 490 |
| Utah International Inc. | 500 | 47 | 23,500 | 30 | 14,750* | 525 |
| | | | 49,050 | | 27,147 | 1,015 |
| *Oil and Gas (10.3 percent)* | | | | | | |
| Continental Oil | 400 | 61 | 24,350 | 33 | 13,158 | 800 |
| Gulf Oil | 1,000 | 21 | 20,500 | 17 | 16,651 | 1,700 |
| | | | 44,850 | | 29,809 | 2,500 |
| *Drugs and Medical (10.5 percent)* | | | | | | |
| Abbot Lab. | 600 | 41 | 24,750 | 37 | 22,253* | 480 |
| Merck | 300 | 69 | 20,775 | 52 | 15,631* | 420 |
| | | | 45,525 | | 37,884 | 900 |
| *Computers (9.7 percent)* | | | | | | |
| International Business Machines | 188 | 224 | 42,159 | 112 | 20,995* | 1,316 |
| *Other Office Equipment (4.3 percent)* | | | | | | |
| Moore Corp. | 400 | 47 | 18,700 | 39 | 15,500* | 480 |
| *Miscellaneous (22.9 percent)* | | | | | | |
| CBS | 500 | 47 | 23,500 | 30 | 14,772* | 830 |
| Disney | 325 | 50 | 16,209 | 3 | 819 | 39 |
| Eastman Kodak | 300 | 106 | 31,838 | 111 | 33,275* | 618 |
| Minnesota Mining and Mfg. | 500 | 56 | 27,750 | 70 | 34,880* | 675 |
| | | | 99,297 | | 83,746 | 2,162 |
| TOTAL EQUITY SECURITIES | | | 433,731 | | 312,000 | 12,845 |
| TOTAL SECURITIES AND CASH | | | 490,672 | | 368,699 | 16,270 |

* More than one cost basis.

large withdrawal made from the Special Grant category.

In terms of the outlook for 1976, the Finance Committee is looking forward to a good year. It is our view that during the next twelve months the economy should continue to recover, the level and quality of profits should continue to improve, the pace of inflation should continue to moderate, and interest rates should continue to stabilize around the lower levels established during the last quarter of 1975. These positive expectations have led the Committee to the judgment that, while equities are clearly not as deeply undervalued as was the case a year ago, the environment for common stock investment nevertheless appears, on balance, quite favorable.

However, it is also the judgment of the Committee, reached only after careful consideration, that an investment program of rolling over high-grade six month commercial paper is, over time, a satisfactory way to protect investable funds from erosion by inflation and that our previous policy of investing only in equities should be replaced by a new policy designed to include a combination of both common stocks and commercial paper.

On the basis of this judgment, the Committee is planning to take advantage of the anticipated strength in the stock market during 1976 to shift one-third of the long term investment account from equities to high-grade commercial paper. Ultimately the Committee intends to have one-half of the portfolio invested in commercial paper, but in view of the likelihood of an up-market trend for common stocks during 1976, the Committee has decided to move only two-thirds of the way toward the target during the next year. In the context of this policy, the Association's investment portfolio will be continuously monitored; and, should events proceed in a way which would justify a shift in emphasis, adjustments will be made in our investment position.

BERYL W. SPRINKEL, *Chairperson*

# AUDITORS' REPORT

*To the Executive Committee of*
*The American Economic Association:*

We have examined the statements of assets and liabilities of THE AMERICAN ECONOMIC ASSOCIATION (a District of Columbia corporation, not for profit) as of December 31, 1975 and 1974, and the related statements of revenues and expenses, changes in general and restricted fund balances and changes in assets and liabilities for the years then ended. Our examination was made in accordance with generally accepted auditing standards, and accordingly included such tests of the accounting records and such other auditing procedures as we considered necessary in the circumstances.

In our opinion, the accompanying statements present fairly the assets and liabilities of The American Economic Association as of December 31, 1975 and 1974, and its revenues and expenses, changes in fund balances and the changes in its assets and liabilities for the years then ended, in conformity with generally accepted accounting principles consistently applied during the periods.

ARTHUR ANDERSEN & CO.

Memphis, Tennessee
February 26, 1976.

THE AMERICAN ECONOMIC ASSOCIATION
STATEMENTS OF ASSETS AND LIABILITIES—DECEMBER 31, 1975 AND 1974

| ASSETS | 1975 | 1974 |
|---|---|---|
| CASH | $121,950 | $ 37,471 |
| INVESTMENTS (Notes 1 and 2): | | |
| Treasury bills and commercial paper, at cost which equals market | 298,000 | 438,000 |
| Corporate stocks, at market (cost of $312,064 in 1975 and $410,753 in 1974) | 433,731 | 391,174 |
| RECEIVABLES, less allowance for doubtful accounts of $1,100 in 1975 and $2,300 in 1974 | 52,312 | 25,552 |
| PREPAID EXPENSES | 4,251 | 1,448 |
| OFFICE FURNITURE AND EQUIPMENT, at cost, less accumulated depreciation of $5,144 in 1975 and $4,161 in 1974 | 7,413 | 8,093 |
| | $917,657 | $901,738 |

| LIABILITIES AND FUND BALANCES | 1975 | 1974 |
|---|---|---|
| ACCOUNTS PAYABLE AND ACCRUED LIABILITIES | $152,136 | $252,710 |
| DEFERRED INCOME (Note 1): | | |
| Life membership dues | 65,480 | 54,774 |
| Other membership dues | 255,992 | 215,298 |
| Subscriptions | 157,818 | 148,961 |
| *Job Openings for Economists* | 11,908 | — |
| | 491,198 | 419,033 |
| ACCRUAL FOR DIRECTORY (Note 1) | 50,000 | — |
| FUND BALANCES: | | |
| Restricted | 139,676 | 307,552 |
| Less—Unrecognized change in market value of investments (Notes 1 and 3) | (7,456) | (68,007) |
| | 132,220 | 239,545 |
| General | 110,535 | 148,381 |
| Less—Unrecognized change in market value of investments (Notes 1 and 3) | (18,432) | (157,931) |
| | 92,103 | (9,550) |
| Total fund balances | 250,211 | 455,933 |
| Less—Unrecognized change in market value of investments (Notes 1 and 3) | (25,888) | (225,938) |
| Net fund balances | 224,323 | 229,995 |
| | $917,657 | $901,738 |

The accompanying notes to financial statements are an integral part of these statements.

THE AMERICAN ECONOMIC ASSOCIATION
STATEMENTS OF REVENUES AND EXPENSES
FOR THE YEARS ENDED DECEMBER 31, 1975 AND 1974

| | 1975 | 1974 |
|---|---|---|
| REVENUES FROM DUES AND ACTIVITIES: | | |
| Membership dues and subscriptions | $380,471 | $350,064 |
| Nonmember subscriptions | 237,191 | 219,299 |
| *Job Openings for Economists* subscriptions | 16,852 | 4,487 |
| Advertising | 63,786 | 61,131 |
| Sale of copies, republications and handbooks | 25,964 | 21,727 |
| Sale of mailing list | 33,397 | 23,540 |
| Annual meeting | 6,906 | 17,427 |
| Sundry | 22,709 | 13,022 |
| | 787,276 | 710,697 |
| INVESTMENT LOSSES (Note 2) | (48,125) | (56,059) |
| Net revenues | 739,151 | 654,638 |
| PUBLICATION EXPENSES: | | |
| *American Economic Review* | 195,627 | 201,082 |
| *Journal of Economic Literature* | 253,407 | 249,359 |
| *Papers and Proceedings* | 50,989 | 51,625 |
| *Directory* (Note 1) | 60,903 | 68,780 |
| *Job Openings for Economists* | 21,788 | 16,472 |
| | 582,714 | 587,318 |
| OPERATING AND ADMINISTRATIVE EXPENSES: | | |
| General and administrative: | | |
| Salaries | 103,996 | 99,401 |
| Rent | 7,854 | 7,962 |
| Other (Exhibit I) | 72,873 | 60,450 |
| Committee | 22,618 | 13,220 |
| Annual meeting | 2,997 | 13,253 |
| Interest and miscellaneous | 1,770 | 1,827 |
| | 212,108 | 196,113 |
| Total expenses | 794,822 | 783,431 |
| REVENUES LESS THAN EXPENSES | $(55,671) | $(128,793) |

The accompanying notes to financial statements and Exhibit I are an integral part of these statements.

THE AMERICAN ECONOMIC ASSOCIATION
STATEMENT OF CHANGES IN GENERAL FUND BALANCE
FOR THE YEARS ENDED DECEMBER 31, 1975 AND 1974

| | Total | Operations | Market Value Adjustments |
|---|---|---|---|
| Balance at January 1, 1974 | $226,610 | $ 62,184 | $164,426 |
| Add—market value increase resulting from inflation (Note 1) | 50,564 | — | 50,564 |
| Deduct—revenues less than expenses | (128,793) | (128,793) | — |
| Balance at December 31, 1974 | 148,381 | (66,609) | 214,990 |
| Add—market value increase resulting from inflation (Note 1) | 17,825 | — | 17,825 |
| Deduct—revenues less than expenses | (55,671) | (55,671) | — |
| Balance at December 31, 1975 | $110,535 | $(122,280) | $232,815 |

The accompanying notes to financial statements are an integral part of this statement.

THE AMERICAN ECONOMIC ASSOCIATION
STATEMENT OF CHANGES IN RESTRICTED FUND BALANCES
FOR THE YEAR ENDED DECEMBER 31, 1974

| | Balance at January 1 | Receipts | Disburse-ments | Allocation of Investment Losses (Note 4) | Balance at December 31 |
|---|---|---|---|---|---|
| The Ford Foundation grants for— | | | | | |
| Economics Institute's orientation program for foreign graduate students of economics | $280,501 | $ 28,710 | $(42,086) | $(11,026) | $256,099 |
| Committee on the status of women in the economics profession | 1,020 | 15,000 | (16,020) | — | — |
| The Alfred P. Sloan Foundation, Chase Manhattan Bank and Ford Foundation grants for increase of educational opportunities for minority students in economics | 23,903 | 57,507 | (58,648) | — | 22,762 |
| Funds reserved by the Association for publication of revised editions of *Graduate Study in Economics*, a guide originally published with funds from a Ford Foundation grant | 6,032 | 62 | — | — | 6,094 |
| The Asia Foundation grant for Asian economists' membership in The American Economic Association and related travel expenses | 1,635 | 2,521 | (3,741) | — | 415 |
| The Carnegie Foundation grant for the committee on the status of women in the economics profession | — | 21,610 | (689) | — | 20,921 |
| The Kazanjian Foundation grant for the committee on economic education | — | 750 | — | — | 750 |
| Sundry | 411 | 100 | — | — | 511 |
| | $313,502 | $126,260 | $(121,184) | $(11,026) | $307,552 |

The accompanying notes to financial statements are an integral part of this statement.

THE AMERICAN ECONOMIC ASSOCIATION
STATEMENT OF CHANGES IN RESTRICTED FUND BALANCES
FOR THE YEAR ENDED DECEMBER 31, 1975

| | Balance at January 1 | Receipts | Disbursements | Allocation of Investment Losses (Note 4) | Balance at December 31 |
|---|---|---|---|---|---|
| The Ford Foundation grant for Economics Institute's orientation program for foreign graduate students of economics... | $256,099 | $ 2,074 | $(135,661) | $(20,508) | $102,004 |
| The Alfred P. Sloan Foundation, Chase Manhattan Bank and Ford Foundation grants for increase of educational opportunities for minority students in economics........................ | 22,762 | 40,551 | (32,560) | — | 30,753 |
| Funds reserved by the Association for publication of revised editions of *Graduate Study in Economics*, a guide originally published with funds from a Ford Foundation grant.................. | 6,094 | — | (6,094) | — | — |
| The Asia Foundation grant for Asian economists' membership in The American Economic Association and related travel expenses.................... | 415 | 2,871 | (1,844) | — | 1,442 |
| The Carnegie Foundation grant for the committee on the status of women in the economics profession........... | 20,921 | — | (16,055) | — | 4,866 |
| The Kazanjian Foundation grant for the committee on economic education.... | 750 | — | (750) | — | — |
| The German Marshall Fund grant for the annual meeting guest speaker........ | — | 1,500 | (1,500) | — | — |
| Sundry............................ | 511 | 100 | — | — | 611 |
| | $307,552 | $47,096 | $(194,464) | $(20,508) | $139,676 |

The accompanying notes to financial statements are an integral part of this statement.

THE AMERICAN ECONOMIC ASSOCIATION
STATEMENTS OF CHANGES IN ASSETS AND LIABILITIES
FOR THE YEARS ENDED DECEMBER 31, 1975 AND 1974

| | 1975 | 1974 |
|---|---|---|
| SOURCE (USE) OF FUNDS: | | |
| Revenues less than expenses | $ (55,671) | $(128,793) |
| Add noncash charges— | | |
| Depreciation | 983 | 737 |
| Directory publication (Note 1) | 50,000 | — |
| Market value adjustments (Note 1) | 80,678 | 90,055 |
| Funds provided (used) by operations | 75,990 | (38,001) |
| (Increase) decrease in— | | |
| Receivables and prepaid expenses | (29,563) | (461) |
| Investments | 97,443 | (53,151) |
| Office furniture and equipment | (303) | (1,924) |
| Increase (decrease) in— | | |
| Accounts payable and accrued liabilities | (100,574) | 151,265 |
| Deferred income | 72,165 | 30,228 |
| Accrual for Directory (Note 1) | — | (80,000) |
| Restricted funds | (167,876) | (5,950) |
| General fund, market value adjustment | 17,825 | 50,564 |
| Unrecognized change in market value of investments | 119,372 | (224,044) |
| Increase (decrease) in cash | $ 84,479 | $(171,474) |

The accompanying notes to financial statements are an integral part of these statements.

THE AMERICAN ECONOMIC ASSOCIATION
NOTES TO FINANCIAL STATEMENTS
FOR THE YEARS ENDED DECEMBER 31, 1975 AND 1974

## (1) Significant Accounting Policies:

*Investments—*

The Association accounts for its investments on a market value basis. Under the method used by the Association to value investments, the change in market value of corporate stocks during the year, after adjusting for an inflation factor (6.4 percent in 1975 and 11.8 percent in 1974), is recognized in income over a three-year period. The changes in market value of investments are allocated to the general and restricted fund balances as described in Notes 3 and 4.

*Accrual for Directory—*

Approximately every three to five years, the Association publishes a directory which lists, among other things, the names and addresses of its membership. This directory was published in 1974 and distributed at no cost to the membership. In order to match more properly the publishing cost of the directory with revenue from membership dues, the Association provides, annually, amounts which in the aggregate will approximate the actual directory expense in the year of publication.

*Deferred Income—*

Payments for membership dues and subscriptions to the various periodicals of the Association are deferred when received; these amounts are then recognized as income as publications are mailed to the members and subscribers.

Life membership dues are also deferred when received; income is recognized over the estimated average life of these members.

## (2) Investments and Investment Income (Losses):

Investment losses recognized in revenues and expenses for the years ended December 31, were as follows:

| | 1975 | 1974 |
|---|---|---|
| Treasury bills and commercial paper— | | |
| Interest | $ 22,449 | $ 23,720 |
| Decline in market value | — | (11,680) |
| | 22,449 | 12,040 |
| Corporate stocks— | | |
| Cash dividends | 15,019 | 16,256 |
| Decline in market value recognized (Note 3) | (113,311) | (117,154) |
| | (98,292) | (100,898) |
| Investment losses allocated to a restricted fund (Note 4) | 27,718 | 32,799 |
| Investment losses included in revenues and expenses | $ (48,125) | $ (56,059) |

## (3) Unrecognized Change in Market Value of Investments:

As described more fully in Note 1, the Association recognizes in income over a three-year period changes in the market value of its corporate stocks. The following summarizes the years in which market value changes in stocks occurred that affect 1974 and 1975 revenues and the amount of these market value increases (declines) that will be recognized in income in future periods.

| Year of Market Value Change | Recognized In Income In | | To Be Recognized In | | Unrecognized Change December 31 | |
|---|---|---|---|---|---|---|
| | 1974 | 1975 | 1976 | 1977 | 1975 | 1974 |
| 1972 | $ 25,069 | $ — | $ — | $ — | $ — | $ — |
| 1973 | (58,509) | (58,509) | — | — | — | (58,509) |
| 1974 | (83,714) | (83,714) | (83,715) | — | (83,715) | (167,429) |
| 1975 | — | 28,912 | 28,913 | 28,914 | 57,827 | — |
| | $(117,154) | $(113,311) | $(54,802) | $28,914 | $(25,888) | $(225,938) |

Included in the above unrecognized changes as of December 31 are declines of $7,456 and $68,007 in 1975 and 1974, respectively, which have been allocated to a restricted fund. The amounts allocated are based on the percentage of the Association's total stock portfolio owned by this restricted fund.

## (4) Restricted Fund:

In 1968, the Association entered into an agreement with the University of Colorado relating to the Ford Foundation grant for the Economics Institute which provides, among other things, that the Association invest a portion of the funds received and allocate any income and market value adjustments therefrom to the restricted fund. In accordance with this agreement, the following adjustments were allocated to the restricted fund:

| | 1975 | 1974 |
|---|---|---|
| Net investment losses (Note 2) | $(27,718) | $(32,799) |
| Market value adjustments arising from inflation | 7,210 | 21,773 |
| | $(20,508) | $(11,026) |

## (5) Retirement Annuity Plan:

Employees of the Association are eligible for participation in a contributory retirement annuity plan. Payments by the Association and participating employees are based on the employee's compensation. Benefit payments are based on the amounts accumulated from such contributions. The total pension expense was $11,910 and $6,798 for 1975 and 1974, respectively.

## (6) The Association:

The American Economic Association files its Federal income tax return as an educational organization, substantially exempt from income tax under section 501(c)(3) of the U.S. Internal Revenue Code. As required by Section 511(a) of this Code, the Association provides for Federal income taxes on certain revenues which are not substantially related to its tax exempt purpose. This "unrelated business income" includes income from advertising and the sale of mailing lists.

The Association has been determined to be an organization which is not a private foundation.

EXHIBIT I
THE AMERICAN ECONOMIC ASSOCIATION
STATEMENTS OF OTHER GENERAL AND
ADMINISTRATIVE EXPENSES
FOR THE YEARS ENDED DECEMBER 31, 1975 AND 1974

| | 1975 | 1974 |
|---|---|---|
| Mailing list file maintenance and periodic mailing expenses | $14,658 | $13,598 |
| Accounting and legal | 7,600 | 7,500 |
| Office supplies | 10,942 | 8,998 |
| Postage | 11,863 | 11,475 |
| Dues and subscriptions | 2,819 | 2,800 |
| Telephone | 2,802 | 3,130 |
| Provision for Federal income taxes (Note 6) | 12,200 | 3,371 |
| Investment counsel and custodian fees | 2,780 | 2,783 |
| President and president-elect expenses | 3,000 | 4,200 |
| Travel and entertainment | 1,732 | 1,327 |
| Depreciation (straight-line method) | 983 | 737 |
| Uncollectible receivables | 51 | 586 |
| Currency exchange charges (credits) | 1,443 | (55) |
| | $72,873 | $60,450 |

# REPORT OF THE MANAGING EDITOR AMERICAN ECONOMIC REVIEW

The number of manuscripts submitted in 1975 was 742, slightly more than last year, and it corresponds to the average rate of the last three years. We printed 112 papers in all, 13 less than last year. The comparative statistics for the last twenty years are shown in Table 1.

TABLE 1—MANUSCRIPTS SUBMITTED AND PUBLISHED, 1955–75

| Year | Submitted | Published | Ratio of Published to Submitted |
|---|---|---|---|
| 1955 | 245 | 41 | .17 |
| 1956 | 242 | 48 | .20 |
| 1957 | 215 | 40 | .19 |
| 1958 | 242 | 46 | .19 |
| 1959 | 279 | 48 | .17 |
| 1960 | 276 | 46 | .17 |
| 1961 | 305 | 47 | .15 |
| 1962 | 273 | 46 | .17 |
| 1963 | 329 | 46 | .14 |
| 1964 | 431 | 67 | .16 |
| 1965 | 420 | 59 | .14 |
| 1966 | 451 | 62 | .14 |
| 1967 | 534 | 94 | .18 |
| 1968 | 637 | 93 | .15 |
| 1969 | 758 | 121 | .16 |
| 1970 | 879 | 120 | .14 |
| 1971 | 813 | 115 | .14 |
| 1972 | 714 | 143 | .20 |
| 1973 | 758 | 111 | .15 |
| 1974 | 723 | 125 | .17 |
| 1975 | 742 | 112 | .15 |

The backlog of accepted papers is slightly smaller than last year. As of November 15, there are 63 accepted papers which have not yet been published. Twenty-five will appear in the March issue, and the remainder in June and September 1976.

As Table 2 indicates we printed 112 papers this year, 53 main articles and 59 shorter papers. The latter include short articles, notes, comments, and replies. The size of the *Review* was reduced to 1068 pages this year from 1146 pages last year. The purpose of the reduction is to control and cut printing costs. This accounts for the smaller number of papers printed, particularly shorter papers.

## Subject Matter of Submitted and Printed Manuscripts

Table 3 presents a distribution of manuscripts classified by subject matter. The most popular fields are microeconomics, labor, welfare theory, macroeconomics, international economics, and monetary theory. These six fields have dominated the distribution for some time.

In addition to submitted papers, the *Review* publishes papers which are com-

TABLE 2—SUMMARY OF CONTENTS, 1974 AND 1975

| | 1974 | | 1975 | |
|---|---|---|---|---|
| | Number | Pages | Number | Pages |
| Articles | 49 | 684 | 53 | 730 |
| Shorter papers, including notes, comments, and replies | 73 | 386 | 59 | 279 |
| Special articles | 3 | 4 | 1 | 2 |
| Dissertations | | 25 | | 22 |
| Announcements and notes | | 47 | | 35 |
| Total | 125 | 1,146 | 113 | 1,068 |

TABLE 3—SUBJECT MATTER DISTRIBUTION OF SUBMITTED AND PUBLISHED MANUSCRIPTS IN 1975

| | Submitted | Published |
|---|---|---|
| General economics and general equilibrium theory | 12 | 1 |
| Microeconomic theory | 87 | 18 |
| Macroeconomic theory | 75 | 9 |
| Welfare theory | 81 | 17 |
| Economic history, history of thought, methodology | 6 | — |
| Economic systems | 30 | 3 |
| Economic growth, development, planning, fluctuations | 30 | 3 |
| Economic statistics | 24 | 6 |
| Monetary and financial theory and institutions | 69 | 8 |
| Fiscal policy and public finance | 29 | 6 |
| International economics | 72 | 11 |
| Administration, business finance | 16 | 4 |
| Industrial organization | 31 | 4 |
| Agriculture, natural resources | 11 | 2 |
| Manpower, labor, population | 112 | 18 |
| Welfare programs, consumer economics, urban and regional economics | 57 | 2 |
| Total | 742 | 112 |

missioned by the editor, or specially chosen for pedagogical purposes.

Under the first heading are commissioned reviews of the annual report of the Council of Economic Advisers. This year there were three reviews: by Harold Carter on the food and agriculture section of the report; by William Poole on the report's analytic content; and a radical critique of the report by Leonard Rapping and James Crotty. It will be recalled that the purpose of commissioned reviews is to stimulate discussion of long-run policy issues, and to satisfy an expressed need for such material by readers who want more policy in their diet. There is no commitment, however, to a permanent annual review of the Council report. Rather than freeze the mold I have decided to experiment with other types of policy discussion. For example, I have commissioned papers on British inflation, to appear later in 1976.

Papers chosen for pedagogical purposes include the Nobel Award addresses delivered by American recipients of this prize.

### Administrative Changes

The financial deficits of the Association in 1974 and 1975 made it necessary to reduce printing and mailing expenses and to increase the Association's revenues. The *Review* was a vehicle for many of these changes. While they will also be discussed in the Treasurer's Report, my report will deal with those which specifically affect the operation of the *Review*.

(a) As indicated earlier, the size of the *Review* has been reduced from 1146 to 1068 net pages. While this required a reduction in the number of papers printed, in the long run control of the length of accepted papers should cut their size and permit a larger number of papers to be printed.

(b) The submission fee for articles and manuscripts is $15 for members and $30 for nonmembers, with the extra fee applicable to membership dues, should the author wish to join. The purpose of this charge is to encourage membership in the Association. A tabulation for the period July 1–October 31, 1975 indicates

TABLE 4—COPIES PRINTED, SIZE, AND COST OF PRINTING AND MAILING IN 1975

| | Copies Printed | Pages | | Cost | | |
|---|---|---|---|---|---|---|
| | | Net | Gross | Issue[b] | Reprints[c] | Total |
| March | 27,500 | 258 | 304 | $ 31,350.37 | $ 425.80 | $ 31,776.17 |
| June | 27,500 | 274 | 296 | 30,072.44 | 571.52 | 30,643.96 |
| September | 27,500 | 258 | 288 | 31,305.89 | 750.00[a] | 32,055.89 |
| December | 27,500 | 278 | 320 | 34,800.00[a] | 200.00[a] | 35,000.00[a] |
| Total | 110,000 | 1,068 | 1,208 | $127,528.70[a] | $1,947.32[a] | $129,476.02[a] |

[a] Estimate.
[b] Includes allocated cost of preparing mailing list.
[c] Costs are reduced by a credit resulting from charges to authors for additional reprints.

that the policy does work. Of 277 papers submitted in that period, 9 authors paid the $30 nonmember fee. It is more significant that 31 authors became new members, or renewed their membership.

The practice of using the submission fee to encourage membership is not a monopoly of our Association. It is also employed by the Royal Economic Society and the Canadian Economics Association.

(c) We have discontinued the distribution of 100 free reprints/offprints to authors. They will be sold to authors if desired.

(d) We have introduced a page charge for articles, shorter papers, notes, comments and replies. This charge of $25 per page is payable by the institution or granting agency supporting the research. Payment of the charge is not a prerequisite for publication, nor are authors expected to pay the charges themselves. Moreover, articles are accepted without prior knowledge of financial support which the author may have. If the author or his institution cannot pay the charge, it is no obstacle to publication.

The purpose of the page charge is to take advantage of the willingness of government sponsoring agencies and foundations to defray the expenses of publishing research in scholarly journals. Page charges are regarded as a legitimate item in the budgets of research grants, and

TABLE 5—ACTUAL AND BUDGETED EXPENDITURES, 1968–76

| | Printing and Mailing | Payments to Contributors | Office Expenses | Total |
|---|---|---|---|---|
| 1968 | $ 92,948 | $2,140 | $29,155 | $124,243 |
| 1969 | 97,183 | 600 | 26,765 | 124,548 |
| 1970 | 111,227 | | 36,336 | 147,564 |
| 1971 | 120,120 | | 43,524 | 163,644 |
| 1972 | 107,196 | | 44,473 | 151,669 |
| 1973 | 117,873 | | 49,121 | 166,994 |
| 1974 | 137,653 | | 58,386 | 196,039 |
| 1975[a] | 165,742 | | 65,911 | 231,653 |
| 1975[b] | 129,476 | | 63,371 | 192,847 |
| 1976[a] | 145,000 | | 69,216 | 214,216 |

[a] Budget.
[b] Actual (estimated).

journals in other disciplines have taken advantage of this source of revenue for some time.

### Expenses—Printing and Mailing

The 1975 costs of printing and mailing the *Review* are shown in Table 4 and for other years in Table 5. The net pages shown include only pages listed in our table of contents. Advertising and index pages then make up the difference between gross and net. The 1975 printing and mailing expenses fell $35,000 short of the budget approved last December. There were a number of reasons: The reduction in net pages (78); the elimination of employment service pages (44); and the failure of a possible increase in printing charges to materialize. The proposed printing and mailing budget will be higher next year, because I expect an increase in mailing charges and in printing costs.

### Office Expenses

Office expenses rose in 1975, because of higher salaries, fringe benefits, and supply costs. I expect a further increase in 1976.

### Board of Editors

Six members of the Board of Editors complete their terms at this time: Bela Balassa, Ann Friedlaender, Stephen Goldfeld, Robert Hall, Leonard Rapping, and Joseph Stiglitz. I wish to thank them for their high standards, hard work, and cooperation.

I should like to express my thanks to the continuing members of the Board of Editors: Eugene Fama, Martin Feldstein, Robert J. Gordon, Bent Hansen, James Melvin, William Nordhaus, Stephen Resnick, Anna Schwartz, Jerome Stein, S. C. Tsiang, Finis Welch, and Marina Whitman.

I am submitting to the Executive Committee the names of six nominees to serve three-year terms until December 31, 1978. They are: Irma Adelman, David Baron, Robert Barro, Laurits Christensen, David Laidler, and Frank Stafford.

The following assisted me this year as editorial consultants: Ernst Berndt, Roger Betancourt, Alan Blinder, Byron Brown, Dennis Epple, Allan Feldman, Donald Frey, William Haley, James Hanson, John Kennan, Edward Lazear, David McNicol, J. David Richardson, Robert Rohr, José Alexandre Scheinkman, Robert S. Smith, John Taylor, Hal Varian, Warren Weher, and Charles Wilson.

I have received valuable proofreading and mathematical assistance from: Clemens F. J. Boonekamp, Marvin Goodfriend, Robert King, Chon Pyo Lee, and Hiroshi Ono. I am grateful to Wilma St. John for her fine work as Assistant Editor, and to Carol Chapin, Editorial Assistant.

In addition to the members of the Board and the editorial consultants, I have sought and received the assistance of a large number of economists during the course of the year. I wish to thank them for their cooperation and high standards in reading and evaluating manuscripts. They have eased the work load that would otherwise fall on the Board of Editors. The following have assisted as referees:

| | |
|---|---|
| I. Adelman | D. P. Baron |
| B. Aghevli | N. S. Barrett |
| N. Aitken | R. J. Barro |
| P. Allen | J. Barron |
| E. Ames | Y. Barzel |
| J. Anderson | R. N. Batra |
| K. J. Arrow | W. J. Baumol |
| O. Ashenfelter | J. R. Behrman |
| C. Azariadis | L. Benham |
| M. Bailey | Y. Ben-Porath |
| R. E. Baldwin | G. Benston |
| E. Baltensperger | U. Ben-Zion |

B. Bergmann
A. Bergson
T. Bergstrom
T. J. Bertrand
S. M. Besen
G. O. Bierwag
A. Blinder
M. K. Block
M. Blume
D. R. Bohi
W. Bomberger
V. Bonomo
K. Borch
M. Boskin
R. Bower
R. Boyer
D. Bradford
G. Brandow
W. H. Branson
R. A. Brecher
F. P. R. Brechling
M. Brennan
R. Britto
W. Brock
C. Brown
J. Brown
G. Brown, Jr.
E. K. Browning
J. M. Buchanan
M. E. Burns
P. Cagan
C. D. Campbell
M. Canes
T. F. Cargill
W. Carleton
J. A. Carlson
F. Casas
M. Chacholiades
P. L. Cheng
B. R. Chiswick
G. C. Chow
L. Christensen
C. Cicchetti
J. Conlisk
J. P. Cooper
J. C. Cox
J. C. Cragg
G. Daly
M. R. Darby
R. d'Arge
E. Davis
K. Davis
R. H. Day
P. Desai
J. DeSalvo
W. E. Diewert
R. K. Diwan
A. K. Dixit
P. Doeringer
F. T. Dolbear, Jr.
G. Douglas
R. Dusansky
G. Eads
N. Edelson
R. Ehrenberg
I. Ehrlich
R. Eisner
B. Ellickson
J. W. Elliott
K. Elzinga
S. Engerman
T. W. Epps
E. W. Erickson
W. Ethier
R. Evenson
L.-S. Fan
A. Feldman
P. J. Feldstein
S. Fischer
A. Fisher
R. Flanagan
F. Flatters
D. Foley
R. Freeman
J. Friedman
I. Friend
V. Fuchs
E. Furubotn
M. Galatin
I. Garfinkel
J. E. Gaummitz
F. Gehrels
H. Genberg
W. E. Gibson
L. Girton
R. Goldfarb
F. L. Golladay
J. P. Gould
E. Gramlich
W. L. Gramm
H. A. J. Green
R. Grieson
J. Griffin
H. Grossman
M. Grossman
M. A. Grove
H. Grubel
J. Hadar
M. Hadjimichalakis
W. J. Haley
R. Hamada
M. Hamburger
D. Hamermesh
O. Hart
W. R. Hart
J. M. Hartwick
J. C. Hause
R. Haveman
J. Heckman
J. M. Heineke
W. P. Heller
D. W. Henderson
J. Henderson
R. Hendler
J. Hirshleifer
W. Holahan
R. Holbrook
C. C. Holt
H. Hori
I. Horowitz
T. Horst
P. Howitt
L. A. Ihnen
Y. M. Ioannides
D. Jaffee
E. James
G. Johnson
M. B. Johnson
T. Johnson
R. W. Jones
P. Joskow
M. Kafoglis
P. Kalman
E. J. Kane
J. H. Kareken
G. G. Kaufman
D. Keesing
J. Kennan
R. A. Kessel
B. Klein
A. K. Klevorick
J. Kmenta
A. V. Kneese
L. A. Kochin
T. Koizumi
R. Korkie
M. B. Krauss
M. E. Kreinin
C. G. Krouse
A. Krueger
P. Kumar
D. Laidler
R. J. Lampman
K. J. Lancaster
L. J. Lau
T. H. Lee
A. Leibowitz
H. Levin
N. Liviatan
R. Lucas
R. E. Lucas
B. T. McCallum
R. McCulloch
R. McKinnon
C. E. McLure, Jr.
M. McManus
G. S. Maddala
S. P. Magee
J. H. Makin
B. Malkiel
J. Marchand
S. A. Marglin
J. Marshall
S. Masters
J. P. Mattila
D. Mayers
T. H. Mayor
D. Meiselman

A. H. Meltzer
L. H. Meyer
R. A. Meyer, Jr.
R. T. Michael
P. Mieszkowski
J. Mingo
H. Minsky
L. Mirman
H. Mohring
S. Morley
J. Muellbauer
A. H. Munnell
R. Muth
E. Nadel
M. I. Nadiri
K. Nagatani
P. Neher
E. J. Nell
C. Nelson
R. Nelson
M. L. Nerlove
D. M. G. Newbery
J. Newhouse
P. K. Newman
A. T. Nichols
J. Niehans
Y. Niho
R. Noll
W. Oakland
W. E. Oates
R. Oaxaca
W. Oi
E. O. Olsen
J. Olson
H. Oniki
J. A. Ordover
D. K. Osborne
J. M. Ostroy
H. Pack
M. Paglin
R. Parks
D. Parsons
D. Patinkin
M. V. Pauly
E. Pazner
M. J. Peck
S. Pejovich
S. Peltzman
J. Pencavel
E. S. Phelps
C. G. Plourde
W. Poole
A. A. L. Powell
F. L. Pryor
D. Purvis
J. Quigley
F. Raines
R. Ramachandran
W. Ramm
A. Razin
D. Roper
S. Rosefielde
S. Rosen
J. Rothenberg
M. Rothschild
S. Rottenberg
M. Rubinstein
W. R. Russell
D. Salkever
P. A. Samuelson
A. Sandmo
A. Santomero
T. Sargent
R. Sato
T. Saving
J. Scadding
L. D. Schall
F. M. Scherer
R. Schmalensee
R. Schuler
C. L. Schultze
M. Schupack
A. Schweinberger
J. Scoville
G. W. Scully
E. Seskin
S. S. Shalit
J. B. Sheahan
W. G. Shepherd
C. D. Siebert
W. Silber
E. Silberberg
N. J. Simler
C. A. Sims
F. Sloan
D. M. Smith
K. Smith
V. L. Smith
E. Smolensky
L. Solmon
R. M. Solow
M. Spence
F. Stafford
S. Star
P. Stephan
J. B. Stephenson
R. M. Stern
W. J. Stull
T. T. Su
J. E. Tanner
P. J. Taubman
M. Taussig
R. L. Teigen
R. W. Thomas
E. A. Thompson
J. Thornton
L. Thurow
T. N. Tideman
T. H. Tietenberg
J. Tobin
R. D. Tollison
J. Triplett
D. Tucker
D. Usher
W. S. Vickrey
G. von Furstenberg
P. Wachtel
M. L. Wachter
N. Wallace
W. G. Waters II
L. Waverman
L. W. Weiss
R. Weiss
J. F. Weston
L. J. White
G. A. Whitmore
C. R. Wichers
J. G. Williamson
J. P. Williamson
R. Willig
D. Wise
D. Wisecarver
J. G. Witte
S. Y. Wu
W. P. Yohe
P. A. Yotopoulos
K. H. Young
R. Zecher
R. Zeckhauser

GEORGE H. BORTS, *Managing Editor*

# REPORT OF THE MANAGING EDITOR JOURNAL OF ECONOMIC LITERATURE

In keeping with this year's shibboleth, "pare costs," this annual report will be somewhat briefer than those of previous years.

At the December 1974 Executive Committee Meeting I was directed to reduce expenditures in any possible way while at the same time maintaining the integrity of the Journal's contents. Members are aware that one way we achieved this goal was to change the shape of the quarterly issues. We reduced the page size slightly but we maintained the same amount of printing on each page. It was the margins that suffered. The monetary savings to the Association were considerable, however. They were achieved because the new size permitted use of a high speed Bookweb press which printed more than six times as fast as the previous press printed. Moreover, the new press used rolls of paper rather than cut sheets. Apparently there are significant economies in omitting the cutting step prior to printing. Finally, our printers, the Kingsport Press, have more sophisticated supplementary machinery (signature folding and arranging) associated with the Bookweb press than they had in the Marioni press complex.

Besides changing the format, we eliminated the Index of Authors of Selected Abstracts section in the June and September issues. It is possible to ferret out where various selected abstracts can be found—instructions are available within each issue of the *Journal* regarding the method to be used—but we have received several strong objections about this decision.

I believe that the editorial office of this Journal has always been run with an eye to cost saving. The trade-off between accuracy and economy means that we proofread our material thoroughly only once. I am told that such weekly news magazines as *Time* proofread their material ten to twelve times; moreover, we use University of Pittsburgh students, rather than professional proofreaders, for this task. During calendar 1975 we continued our usual monitoring of intra-office productivity; I do not think it correct to report that we increased our cost consciousness, but the Associate Editor did reallocate her time (involving extra work) to make sure that the "learning curve" associated with annotators and classifiers was steepened. Moreover, the Assistant Editor (again by working longer hours) took increased care to steepen the learning curve of student personnel involved in proofreading. The Assistant Editor and the Associate Editor also increased their duties in order to speed up the production of the annual indexes (details about this will come later in this report). Our secretarial staff (Mrs. Lyndis Rankin, Mrs. Margaret Yanchosek, and Miss June Cox) plus a regular but temporary secretary (Miss Ninamary Sawich) worked hard and long in order to process not only the four quarterly issues but also the material for the forthcoming annual indexes.

Table 1, JEL-Index Expenditures 1971 through 1975, does not perfectly reflect the productivity picture. Our office expenditures were less than was authorized, even though we processed a great deal of material for the aforementioned annual indexes. As of this time of writing, it seems likely that we will have spent on

TABLE 1—*JEL*-INDEX EXPENDITURES 1971 THROUGH 1975[a]

| | *JEL* Printing[b] | *JEL*-Index Editorial, Bibliographical, Payments to Contributors, etc. | Total |
|---|---|---|---|
| 1971 | $115,127 | $ 74,284 | $189,410 |
| 1972 | 118,442 | 85,083 | 203,525 |
| 1973 | 134,418 | 87,032 | 221,450 |
| 1974 | 162,111 | 87,248 | 249,359 |
| 1975 | 170,000 | 103,778 | 273,778 |
| Budget estimated expenditures | 159,200 | 98,785 | 257,985 |

[a] Figures for 1971–74 are coordinated with the auditor's figures.
[b] Includes mailing, addressing, and gross cost of reprints.

printing, including mailing, addressing, and gross cost of reprints, less than we spent in 1974 and considerably less than we were authorized to spend in 1975 (before the decision to change to the Bookweb press had been made).

I have mentioned the annual indexes. Unlike the quarterly journal where journal articles are classified according to a 3-digit 100-pigeon hole system, the annual indexes (index for articles appearing in a particular calendar year) are classified according to a compatible 4-digit 300-pigeon hole system. These annual indexes include not only the material found in the quarterly issues of the *Journal of Economic Literature* but also essays appearing in *Festschriften*, collections of reprinted works, and the like. For some years we also have selected government documents, including particularly testimony before Congressional Committees.

The 1970 *Index* which has been at the printers since the fall of 1974, is now finally in page proof. The 1971 *Index* is in galleys. We anticipate now that the

TABLE 2—QUANTITATIVE ANALYSIS OF CONTENTS, *JEL*, 1971 THROUGH 1975

(Number of pages in brackets)

| | 1971 | | 1972 | | 1973 | | 1974 | | 1975 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | No. | Pages | No. | Pages | No. | Pages | No. | Pages | No. | Pages |
| Survey articles | 1 | [37] | 3 | [78] | 4 | [144] | 3 | [102] | 3 | [119] |
| Essays on subfields | 5 | [105] | 5 | [129] | 2 | [28] | 5 | [99] | 5 | [100] |
| Review articles | 5 | [40] | 5 | [39] | — | — | 1 | [5] | — | — |
| Articles about economic literature | 2 | [32] | 1 | [19] | 2 | [38] | — | — | 1 | [11] |
| Communications | 7 | [24] | 4 | [13] | 10 | [26] | 13 | [71] | 12 | [36] |
| Books annotated | 1,279 | [206] | 1,209 | [257] | 1,214 | [239] | 1,211 | [229] | 1,203 | [223] |
| Books reviewed | 172 | [249] | 160 | [241] | 175 | [259] | 168 | [239] | 183 | [282] |
| Journal issues listed and indexed | 835 | [130] | 849 | [140] | 1,011 | [185] | 986 | [180] | 908 | [177] |
| No. of individual articles | | | 5,387 | | 7,218 | | 7,360 | | 6,788 | |
| Subject index of journal articles | — | [251] | — | [263] | — | [357] | — | [338] | — | [349] |
| Abstracts of articles | 1,426 | [227] | 1,415 | [313] | 1,906 | [407] | 1,645 | [312] | 1,637 | [331] |
| Total pages[a] | | [1,479] | | [1,572] | | [1,748] | | [1,671] | | [1,700] |

[a] Includes, in addition to listed pages, classification systems, table of contents, indices, journal subscription information, etc.

Kingsport Press has mastered the production problems, its end of the printing process will proceed more rapidly, and the other volumes that are ready for the printer, 1969 and 1972, will also appear in 1976. The speed with which further volumes can be published is dependent on the supply of part-time labor plus the ability of the editorial staff to train and supervise the additional employees and of the entire full-time staff to process the material. However, by 1977 index production should be as current as technically feasible; that is, the final receipt, classification, and publication preparation of all journals and books for a calendar year should be completed within a two-year period.

Table 2 illustrates the subject matter covered in the 1975 volume. We fell short of four survey articles by one. On the other hand, we published about the number of essays and review articles that we normally anticipate. We did not publish a section in the September issue listing journal information. For this omission, there were two reasons; one was the omnipresent concern regarding cost paring, the other was a realization that during periods of price inflation, price information regarding subscriptions tends to become quickly obsolete.

Perceptive readers will note that we have somewhat fewer journal issues listed and indexed in the 1975 volume (compared to the two previous years). For the most part this reduction has been the result of completion of the "catching-up" process in journal listing. During the previous two years we had improved our ability to ensure orderly receipt of the necessary materials from our 220 journals. We have also speeded up the publication process—we classify, list, and index with less time-lag in our editorial office. But if we have processed somewhat fewer journal issues, we have at the same time increased the ratio of abstracts of articles to total listings (in 1974 about 22.4 percent of the listed articles were abstracted; in 1975 about 24.1 percent).

We anticipate that the contents of the 1976 journal will include four survey articles (all of which are in fairly advanced stages of preparation), four articles on an individual scholar's reaction to literature emerging in his (her) field, and possibly one essay on the role of economics journals within the profession. Such has been our objective in the past; I think in 1976 we should have no trouble in meeting it.

Tables 3 and 4 are self-explanatory. Table 3 indicates that the objective of eschewing difficult articles has been somewhat better met this year than in previous years. Table 4 reveals (if compared to the similar tables in previous years) that we have broadened the spectrum (number of categories treated).

The Chancellor and Dean of the Faculty of Arts and Sciences of the University of Pittsburgh have once again allocated some University of Pittsburgh support to the *Journal*. Their willingness to do so, particularly in a period of tight budgets, illustrates an understanding of and a devotion to scholarly work in the economics discipline. I have regularly thanked them privately and take this opportunity again to do so publicly.

TABLE 3—CLASSIFICATION BY TECHNICAL DIFFICULTY, 1969–75 (incl.)

| | Surveys | Creative Curmudgeon Articles | Others[a] | Total |
|---|---|---|---|---|
| Most difficult | 7 | 8 | 3 | 18 |
| Some difficulty | 11 | 14 | 13 | 38 |
| Not difficult | 4 | 4 | 24 | 32 |
| Totals | 22 | 26 | 40 | 88 |

[a] Review articles or books and general essays on all literature; excludes very short communications.

TABLE 4—CLASSIFICATION BY SUBJECT, 1969–75 INCL.

| | Commissioned Survey | Creative Curmudgeon Essays | Other[a] | Total |
|---|---|---|---|---|
| 01 General | 1 | — | 5 | 6 |
| 02 Theory | 5 | 8 | 4 | 17 |
| 03 Thought (Methodology) | — | 2 | 17 | 19 |
| 04 Economic History | — | — | 2 | 2 |
| 05 Comparative Systems | 1 | 2 | 1 | 4 |
| 11–12 Growth & Development | 1 | 5 | — | 6 |
| 21–22 Econometric, Statistical Theory, Statistics | 2 | — | 1 | 3 |
| 31 Monetary Economics | — | 2 | 1 | 3 |
| 32 Fiscal Economics | 1 | 2 | 1 | 4 |
| 40–44 International Economics | 2 | 2 | 7 | 11 |
| 50 Managerial Economics | — | 1 | — | 1 |
| 60 Industrial Organization, Industrial Regulation | 1 | — | — | 1 |
| 70 Agricultural Economics | 1 | — | — | 1 |
| 80 Labor Economics | 2 | 2 | — | 4 |
| 90 Applied Welfare Economics, Regional Economics | 5 | — | 1 | 6 |
| Totals | 22 | 26 | 40 | 88 |

[a] Review articles on books, general essays on all literature.

Three members of the Board of Editors have completed their terms. They have been splendid colleagues, and I wish to convey to them publicly (although I have privately) my great appreciation. Professor Martin Bronfenbrenner (Duke University), Professor Anne O. Krueger (University of Minnesota), and Professor Allan H. Meltzer (Carnegie-Mellon University) have been the quintessence of what every managing editor wants. The other members of my Board have also been superb colleagues. I look forward to my continuing association with them. I have nominated several people to replace the departing three; it is, however, premature for me to announce the names in this report (the procedure is for me to nominate but for the Executive Committee, which meets after this report has been prepared, to confirm my nominations).

I wish also to thank the following economists (plus three who chose to remain anonymous) for advice and assistance in the commissioning, refereeing, and revising articles:

| | |
|---|---|
| Robert Adams | Michio Morishima |
| Albert Ando | Chandler Morse |
| Harvey Averch | Leon Moses |
| Robin Barlow | Richard A. Musgrave |
| Harold J. Barnett | Alan Peacock |
| William Baumol | Joseph A. Pechman |
| Sanford V. Berg | Wilfred Prest |
| Barbara Bergmann | Barbara Reagan |
| Abram Bergson | Harry Richardson |
| James Buchanan | Lawrence S. Ritter |
| Richard Caves | Eugene Rotwein |
| A. W. Coats | Paul Samuelson |
| Richard M. Cyert | Isabel V. Sawhill |
| Horace J. De Podwin | Martin Segal |
| George Feiwel | George L. S. Shackle |
| Martin S. Feldstein | Martin Shubik |
| Margaret S. Gordon | David Starrett |
| Arnold Harberger | Marina v. N. Whitman |
| Alfred Kahn | Gordon C. Winston |
| Michael C. Lovell | Basil Yamey |

MARK PERLMAN, *Managing Editor*

# REPORT OF THE DIRECTOR
# JOB OPENINGS FOR ECONOMISTS

During 1975, 567 employers advertised 1,383 openings in *Job Openings for Economists* (*JOE*). Of these, 1,243 were jobs not previously listed. About 66 percent of the new positions were classified as academic and about 34 percent were nonacademic. As expected the December issue carried the largest number of listings. It contained 30 percent of all jobs listed during the year and 35 percent of the new ones. Table 1 shows the total listings, total jobs, new listings, and new jobs by type for each of the 1975 issues.

Universities with graduate programs and 4-year colleges were the major advertisers of jobs—43 and 40 percent of employers respectively. Nonacademic employers accounted for approximately 17 percent. Table 2 shows the number of employers by type for each issue.

The field of specialization most in demand was general economic theory (020), followed by econometrics (210). Generalists with a strong background in mathematics and statistics who can teach one applied area appear to be the type of economists that employers are seeking. The specific applied area seems to be of secondary importance. Table 3 lists the number of citations for each issue during 1975 for each of the broader field classifications. General economic theory led with 22 percent of the total number of citations, followed by business administration, finance, marketing and accounting (13 percent), monetary and fiscal (11

TABLE 1—JOB LISTINGS FOR 1975

| Issue | Total Listings | Total Jobs | New Listings | New Jobs |
|---|---|---|---|---|
| Academic | | | | |
| February | 115 | 253 | 73 | 115 |
| April | 88 | 154 | 65 | 100 |
| June | 59 | 100 | 51 | 80 |
| August | 38 | 70 | 34 | 59 |
| October | 75 | 150 | 64 | 126 |
| December | 160 | 369 | 147 | 340 |
| Subtotals | 535 | 1,096 | 434 | 820 |
| Nonacademic | | | | |
| February | 16 | 80 | 15 | 78 |
| April | 12 | 59 | 8 | 43 |
| June | 18 | 105 | 15 | 90 |
| August | 18 | 92 | 15 | 77 |
| October | 19 | 58 | 14 | 45 |
| December | 27 | 110 | 23 | 90 |
| Subtotals | 110 | 504 | 90 | 423 |
| Totals: Academic and nonacademic | 645 | 1,600 | 524 | 1,243 |

TABLE 2—NUMBER AND TYPES OF EMPLOYERS LISTING POSITIONS IN *JOE* DURING 1975

| Issue | 4-Year Colleges | Universities with Graduate Programs | Junior Colleges | Federal Government | Federal Reserve Banks | Other | Total |
|---|---|---|---|---|---|---|---|
| February | 52 | 63 | | | | 16 | 131 |
| April | 47 | 40 | 1 | 2 | | 10 | 100 |
| June | 34 | 23 | 2 | | | 18 | 77 |
| August | 16 | 22 | | 3 | 1 | 14 | 56 |
| October | 34 | 41 | | 5 | 1 | 13 | 94 |
| December | 72 | 88 | | 10 | 1 | 16 | 187 |
| Totals | 255 | 277 | 3 | 20 | 3 | 87 | 645 |

percent), welfare and urban (10 percent), and econometrics and statistics (10 percent).

*JOE* is now virtually self-supporting. The estimated 1975 deficit is $451 (see Table 4 for the details of expenses and revenues). Total revenues are estimated to be $17,615 and total expenditures $18,066. The costs include an allocation of salaries and fringe benefits from the budget of the Secretary-Treasurer's office equivalent to one full-time staff employee, but do not include an allocation for overhead (utilities, space, telephone, etc.). The Secretary-Treasurer agrees that an allocation of overhead is not warranted.

TABLE 3—FIELDS OF SPECIALIZATION CITED

| Field[a] | February 1975 | April 1975 | June 1975 | August 1975 | October 1975 | December 1975 | Totals |
|---|---|---|---|---|---|---|---|
| General economic theory 000 | 63 | 61 | 46 | 28 | 54 | 147 | 399 |
| Growth and development 100 | 27 | 16 | 10 | 11 | 14 | 43 | 121 |
| Econometrics and statistics 200 | 40 | 21 | 17 | 22 | 25 | 56 | 181 |
| Monetary and fiscal 300 | 46 | 24 | 16 | 20 | 30 | 69 | 205 |
| International economics 400 | 13 | 14 | 8 | 11 | 8 | 38 | 92 |
| Bus. ad., finance, marketing and acctg. 500 | 51 | 39 | 23 | 19 | 32 | 80 | 244 |
| Industrial organization 600 | 30 | 13 | 17 | 13 | 11 | 38 | 122 |
| Agriculture and natural resources 700 | 17 | 9 | 10 | 8 | 9 | 24 | 77 |
| Labor 800 | 27 | 12 | 14 | 4 | 16 | 36 | 109 |
| Welfare and urban 900 | 59 | 27 | 16 | 14 | 23 | 50 | 189 |
| Related disciplines A00 | 1 | 1 | 1 | 1 | 0 | 3 | 7 |
| Administrative positions B00 | 8 | 10 | 8 | 6 | 12 | 18 | 62 |
| Totals | 382 | 247 | 186 | 157 | 234 | 602 | 1,808 |

[a] Fields of specialization codes are from the *Journal of Economic Literature*.

TABLE 4—PROJECTED REVENUES AND EXPENSES FOR 1975 AND 1976

| | 1975 | 1976 |
|---|---|---|
| Revenues: | | |
| Subscriptions | $17,167 | $19,000 |
| Mailing List | 361 | 300 |
| Advertising | 87 | 50 |
| Total | $17,615 | $19,350 |
| Expenses: | | |
| Addressing and Postage | $ 4,171 | $ 5,420 |
| Printing | 3,857 | 4,000 |
| Computer Time | 1,145 | 1,200 |
| Allocated Salaries | 8,893 | 9,230 |
| Total | $18,066 | $19,850 |
| Deficit | $ (451) | $ (500) |

The projected deficit for 1976 is $500. Revenues are estimated to be $19,350, and expenses to be $19,850. Subscription revenues are estimated conservatively. The projection is based on the current number of subscribers (1,662 at the end of November 1975), and assumes no change during the coming year. Experience indicates that the number will increase. From December 1974 to August 1975, subscribers increased from 2,182 to 2,752. A similar pattern of growth will probably occur in 1976.

Because *JOE* is now essentially self-supporting, I do not recommend an increase in subscription rates or the initiation of a charge for listing at this time. Continued close scrutiny of costs should be sufficient to maintain a balanced budget for at least one more year.

All the comments we have received indicate that *JOE* performs a valuable service for the Association, and that it is an important addition to the flow of information about the job market. According to a survey of placement officers for Ph.D. granting departments, *JOE* ranked third (out of eight) as the original source of information about jobs eventually accepted by their students. It ranked below informal contacts and departmental placement services, but above the placement facilities at the annual meeting.

We expect an expanding role for *JOE* as a source of job information for economists.

C. ELTON HINSHAW, *Director*

# REPORT OF THE COMMITTEE ON THE STATUS OF WOMEN IN THE ECONOMICS PROFESSION

Four years ago the American Economic Association (*AEA*) took a firm stand on the need (a) to support and facilitate equality of opportunity for women economists in all aspects of economists' professional activities and (b) to help eradicate any institutional or personal discrimination against women economists. In May of 1972 the Committee on the Status of Women in the Economics Profession (*CSWEP*) came into being as an *ad hoc* committee of the Association, and in March of 1974, *CSWEP* was made a standing committee of *AEA* in recognition of the fact that the problems addressed by *CSWEP* require long-run efforts. Throughout the four years, it has been a very active committee. This year the committee size was reduced from 10 to 6 members as an economy measure, and we have rotated committee membership. The reduction in size has made it difficult to have as wide geographic representation in any one year as would be desirable and has increased the burden of members on the smaller committee. I want to thank the committee members who have worked so diligently in 1975 to carry out the mandate of the Association.

I want to thank the Carnegie Foundation for its $21,600 grant to us to plan and execute a national, working research conference on occupational segregation and to disseminate the results. I also want to thank the German Marshall Fund of the United States for a S1,500 grant to enable us to enrich the program at the *AEA* annual meeting this year by bringing Baroness Nancy Seear from the London School of Economics to speak on her research on employment conditions in the United Kingdom with particular emphasis on the status of women workers.

The number of women economists identified by *CSWEP* and associated with our activities has continued to grow. By last year, we had identified nearly 1,400 women economists through questionnaires to all colleges and universities with economics departments, letters to government agencies and businesses, and information from the informal network. Then, in November 1974 the *AEA Directory of Members* came out. We requested a list of women from that source. A comparison of the two lists added 313 more names. We contacted these women, and only one or two asked not to be added to our roster of associate members.

We also obtained new names from the 1975–76 Universal Academic· Questionnaire, wherein each fall new hires are reported as well as other data for an analysis of the academic labor market. Our analysis of the data by sex is in the final section of this report.

The total number of women economists now identified is about 1,800. Of these, 20 percent are in government and business (12 percent and 8 percent, respectively). We know that for the moment we have nearly all women members of the American Economic Association, plus an additional group of women economists who do not belong to *AEA*. About 20 percent of those on our roster do not belong to *AEA*. We suspect that there are still more women economists in the government and in business whom we have not identified. We are working on this. In Sep-

tember we met with members of the National Economists Club in Washington, D.C. to report on our research findings and to identify more women economists who would like to register with us. We encourage growth of local subgroups of women economists. Such groups identify more of the women economists, and increase *AEA* membership.

During the past year *CSWEP* has continued to improve the operation of the market for economists, has increased the effective supply of women economists, and has added to the research information on the status of women.

### I. CSWEP Newsletter

Four *CSWEP Newsletters* have been written during 1975 and sent to 1,800 women economists. The newsletter improves the working of the market by listing job openings for economists submitted to us by prospective employers. This activity fills a need most appreciated by agency and department heads. The *CSWEP Newsletter* lists requests for articles, conference and program plans and participants, grant and fellowship opportunities, and notices of regional activities for women economists. It is also a mechanism by which the Committee can ask the reader for help in various *AEA* projects. The newsletter clearly has helped to widen the informal network and to support our women colleagues who often have felt isolated in our profession.

### II. Roster

The roster of women economists developed and computerized by this Committee has been maintained this year. It has been updated by sending to all women economists listed copies of the material they had previously supplied us and asking for the most recent information on their area of specialization, highest degree in economics, professional grade or rank, and address. The roster has been used to provide prospective employers with a list of women economists who meet specified criteria such as particular field of specialization, degree, and rank. The prospective employer may then write to these economists, ask them to supply detailed career data, and finally follow up with interviews if desired. This service is being provided for a fee to help defray committee costs, and its use is growing.

### III. Research Function

An explicit charge to collect and analyze data relevant to the status of women economists and to further the theoretical and applied research related to the status of women in general helps set the work of the standing committee of *AEA* apart from caucuses in some other professional associations. This research dimension is one that *CSWEP* feels is vital for professional associations to undertake to help build a solid foundation for policy prescriptions related to reappraisal of women's place in society. I want to mention in particular this year work on four of our research projects.

*Workshop Conference on Occupational Segregation.* This was jointly sponsored in May 1975 by *CSWEP* and the Center on Research on Women in Higher Education and the Professions at Wellesley College and funded by the Carnegie Corporation of New York. The idea to hold a workshop conference on this topic evolved from *CSWEP*s desire to make a research contribution to International Women's Year on a topic central to an improvement of the status of women in American society.

The papers which we commissioned for that conference plus some additional analysis will be published by the University of Chicago Press in a book supplement to

the new journal, *Signs,* in about March 1976, and subsequently as a separate book *Women in the Workplace, An Analysis of Occupational Segregation,* Martha Blaxall and Barbara B. Reagan (eds.). This will fulfill our promise to disseminate the research results beyond the small working group at the conference. We thank the Carnegie Foundation for making this effort financially possible. Clearly women's economic status in a macro sense is of critical importance to women economists, as well as to the nation. An improvement in women's economic status cannot be realized without a diminution of occupational segregation and a better understanding of the forces behind it. Even within our profession of economics, some fields of specialization are deemed more appropriate for men and others for women. Sexism in occupational segregation is pervasive down to micro levels.

*1974–75 Survey of Economics.* With the development of a roster of women economists and the new *AEA Directory of Members,* it became possible to conduct a survey of both male and female economists to compare educational attainment, employment history, and salaries. Preliminary results from the 1974–75 Survey of Economists were reported in the May 1975 *American Economic Review.* Subsequently, data collection from the men economists in the paired sample was completed. *CSWEP* is now in process of completing the study of 1,240 male and female economists who were paired by selecting a male economist from the same graduate class of a woman economist at the school from which she got her highest degree. The major finding so far is that salaries of women economists are substantially below those of their male counterparts even when their educational attainments and work histories are similar. Excluding the few who were not working, the average income for the women in 1973-74 (7 years after their Ph.D.s) was $15,310. For men in the comparable group of Ph.D.s who were working, it was $19,025 or 24 percent higher. Only 16 percent of the women made more than $25,000, whereas 30 percent of the men economists in the matched group did. Untangling the factors that might explain this income difference is obviously the next step. It clearly isn't difference in formal education in economics—either in amount or quality of school attended. Discrimination against women is morally and legally untenable.

The Committee feels that increased efforts on its part should be directed to reducing the existing discrepancies. The pending exemption of universities and colleges from the affirmative action regulations makes an increased effort on the part of the Committee and the Association even more imperative.

Other findings from the paired sample are of interest to those involved in training economists. For example, the length of time to complete the Ph.D. degree is not significantly different between women and men—15 percent of women and 15 percent of men finished in 3 years or less, 45 percent of the women and again 45 percent of the men finished in 4 or 5 years; 26 percent of the women and 24 percent of the men took 6 or 7 years; 8 percent of the women and 10 percent of the men took 8 or 9 years; and 7 percent of the women and 6 percent of the men took 10 or more years to finish. Considering the averages, women took 5.6 years and men took 5.5 years to complete the Ph. D.

Marriage and increasing family size, however, may contribute to a small delay in starting a Ph.D. The average number of years between the B.A. degree and the Ph.D. in economics was about 8 years for men but 9 years for women. A counterbalancing factor was that the women had

TABLE 1—DEGREES GRANTED IN ECONOMICS BY TYPE OF DEPARTMENT AND SEX, 1974–75

(Departments reporting by December 10, 1975 on 1975–76 Universal Academic Questionnaire)

| Degrees Granted in 1974–75 | All Departments | Ph.D. Chairman's Group | Ph.D. Other Departments | M.A. | B.A. |
|---|---|---|---|---|---|
| Number of departments reporting | 375 | 38 | 27 | 38 | 272 |
| Ph.D., number | 501 | 398 | 103 | — | — |
| Percent women | 10.6 | 10.6 | 10.7 | — | — |
| M.A., number | 1,090 | 544 | 282 | 264 | — |
| Percent women | 18.3 | 18.9 | 20.2 | 15.1 | — |
| B.A., number | 8,402 | 2,219 | 865 | 710 | 4,608 |
| Percent women | 22.2 | 23.6 | 16.3 | 17.6 | 23.3 |
| Other degrees from economics departments, number | 69 | 50 | 2 | 8 | 9 |
| Percent women | 20.3 | 10.0 | 0 | 0 | 100.0 |

gotten their B.A. degrees at an age a year younger than the men had.

*Session at AEA Meetings*. At the 1975 annual *AEA* meetings, a session on the economic status of women in the United States and in the United Kingdom was presented. Two papers surveying current research on employment of women were commissioned, and the policy implications of the research were discussed.

*Academic Labor Market, 1974–75*. For the first time in four years, the proportion of women among those receiving undergraduate degrees and M.A. degrees in economics increased.[1] In 1974–75, of the B.A. degrees in economics 22 percent were earned by women. (See Table 1.) In the two previous years, women comprised only 16 percent of the economics undergraduate majors. The proportion of M.A. degrees earned by women in 1974–75 was 18 percent, compared with 14 percent the two previous years. This increase occurred in those departments that offer both Ph.D. and M.A. degrees. The proportion of Ph.D. degrees earned by women was 11 percent in 1974–75, compared with only 8 percent last year, and with 12 percent the two previous years.[2]

Nearly 75 percent of the full-time Ph.D. students received financial aid in the fall of 1975, and more than half of the full-time M.A. students received financial aid (Table 2). (The latter proportion is down somewhat from last year.) Women students at the Ph.D. level received financial aid approximately in proportion to their numbers, whereas women at the M.A. level received financial aid somewhat more often than proportional to their numbers.

[1] In 1975–76 for the fourth year, data related to supply of economists and academic demand for them are available from a survey of academic departments of economics. The data from the 1975–76 Universal Academic Questionnaires have been collected under the direction of C. Elton Hinshaw of *AEA*, and the data classified by sex are analyzed here. The questions asked in the 1975–76 survey are for the most part comparable to the data published in the *CSWEP* Report last year in the May *American Economic Review*. The number of departments which had reported in time for this analysis is 375 this year, but was only 311 last year. Not all of the departments who reported last year reported again this year. Thus, comparisons of absolute numbers must be made with care. Percentages are more comparable, although, of course, they are subject to sampling error.

[2] The data in the first two years related to students enrolled in the programs, while in the last two years the data refer to degrees received.

TABLE 2—NUMBER OF FULL-TIME "ON-CAMPUS" GRADUATE STUDENTS REGISTERED FALL, 1975 AND TYPE OF FINANCIAL AID, BY TYPE OF DEPARTMENT AND BY SEX

(Departments reporting by December 10, 1975 on 1975–76 Universal Academic Questionnaire)

| Type of Department, Degree Sought, and Sex | Total | Full-Time Students Receiving Financial Aid: Tuition Only | Stipend Only | Tuition and Stipend | No Aid |
|---|---|---|---|---|---|
| All departments: | | | | | |
| Ph.D. students, number | 2,237 | 158 | 469 | 1,039 | 571 |
| Female as percent of total | 12.9 | 14.6 | 12.2 | 13.5 | 12.1 |
| M.A. students, number | 1,227 | 54 | 196 | 397 | 580 |
| Female as percent of total | 17.8 | 29.6 | 21.4 | 19.9 | 14.3 |
| Chairman's group: | | | | | |
| Ph.D. students, number | 1,808 | 107 | 362 | 873 | 466 |
| Female as percent of total | 13.4 | 18.7 | 13.0 | 13.6 | 12.0 |
| M.A. students, number | 530 | 13 | 75 | 196 | 246 |
| Female as percent of total | 17.4 | 38.5 | 33.3 | 13.8 | 14.2 |
| Ph.D., other departments: | | | | | |
| Ph.D. students, number | 429 | 51 | 107 | 166 | 105 |
| Female as percent of total | 11.0 | 5.8 | 9.3 | 12.6 | 12.4 |
| M.A. students, number | 423 | 32 | 96 | 120 | 175 |
| Female as percent of total | 18.0 | 28.1 | 13.5 | 22.5 | 15.4 |
| M.A. departments: | | | | | |
| M.A. students, number | 274 | 9 | 25 | 81 | 159 |
| Female as percent of total | 18.6 | [a] | 16.0 | 30.9 | 13.2 |

[a] Percentage not shown when fewer than 10 in cell.

This was particularly true in departments in the Chairman's Group in 1975. The increase in financial aid in the fall of 1975 to women students at the M.A. level should help perpetuate the increase in the proportion of women receiving M.A.s from 1973–74 to 1974–75 noted above.

Hopefully, next year, the increase in women receiving Ph.D.s will also continue to surpass the previous levels. However, the selected data available now do not suggest that this will occur. The proportion of women receiving M.A. degrees in 1973–74 who entered doctoral programs in 1974–75 was only 12 percent compared with 20 percent of the men (Table 3). In the previous year's survey the proportion of women with M.A. degrees from the year before who entered the Ph.D. program was 24 percent. This was nearly as high as the proportion among the men. Thus, the decrease in doctoral students entering from the M.A. level occurred for both sexes, but more often among women.

Women who received their Ph.D. degrees in economics in 1974–75 found jobs in academic institutions less often than was true for the class of the previous year. While 57 percent of the men Ph.D.s in the 1974–75 class found employment as economists in academia, only 33 percent of the women did.[3] In the previous year, 68 percent of the women with new Ph.D.s were employed in academic institutions. The decline in the academic market thus

[3] If the women whose current occupation was not reported are distributed, the proportion rises to over 40 percent.

TABLE 3—1975-76 EMPLOYMENT OF 1974-75 GRADUATES IN ECONOMICS BY LEVEL OF DEGREE, ALL DEPARTMENTS AND CHAIRMAN'S GROUP, BY SEX

(Departments reporting by December 10, 1975, on 1975-76 Universal Academic Questionnaire)

| Type of Department and Kind of Employment | Ph.D.[a] | | M.A. | |
|---|---|---|---|---|
| | Male | Female | Male | Female |
| All departments: | | | | |
| Number of 1974-75 graduates | 620 | 123 | 605 | 116 |
| Percent | 100.0 | 100.0 | 100.0 | 100.0 |
| Percent employed as economist in United States: | | | | |
| Educational institution | 57.2 | 33.3 | 6.8 | 10.3 |
| Business or industry | 4.2 | 2.4 | 10.9 | 9.5 |
| Federal government | 9.8 | 4.9 | 6.4 | 5.2 |
| State/local government | 3.4 | 3.3 | 7.1 | 7.8 |
| Banking or finance | 2.6 | 1.6 | 5.0 | 3.4 |
| Consulting/research | 6.3 | 4.1 | 3.3 | 2.6 |
| Percent employed outside United States | 6.6 | 21.1 | 18.5 | 19.8 |
| Percent unemployed, seeking work | 2.1 | 2.4 | 2.3 | 0.9 |
| Percent not in labor force: | 3.4 | 4.1 | 22.7 | 19.0 |
| Entered postdoctoral program or Ph.D. program | 0 | 0 | 20.2 | 12.1 |
| Other | 3.4 | 4.1 | 2.5 | 6.9 |
| Percent not known | 4.4 | 22.8 | 17.0 | 21.5 |
| Chairman's Group: | | | | |
| Number of 1974-75 graduates | 460 | 70 | 250 | 58 |
| Percent | 100.0 | 100.0 | 100.0 | 100.0 |
| Percent employed as economist in United States: | | | | |
| Educational institution | 56.8 | 40.1 | 4.0 | 8.6 |
| Business or industry | 3.7 | 2.8 | 13.2 | 13.8 |
| Federal government | 9.3 | 7.1 | 7.2 | 1.7 |
| State/local government | 2.8 | 1.4 | 5.6 | 1.7 |
| Banking or finance | 2.6 | 4.2 | 7.6 | 6.9 |
| Consulting/research | 4.1 | 8.6 | 4.0 | 1.7 |
| Percent employed outside United States | 11.8 | 25.8 | 17.2 | 17.2 |
| Percent unemployed, seeking work | 1.5 | 2.9 | 1.2 | 0 |
| Percent not in labor force: | 2.1 | 0 | 25.6 | 32.7 |
| Entered postdoctoral or Ph.D. program | 0.4 | 0 | 23.2 | 20.7 |
| Other | 1.7 | 0 | 2.4 | 12.1 |
| Percent not known | 5.3 | 7.1 | 14.4 | 15.7 |

[a] Includes graduate students who have not completed their dissertations, if they entered the labor market seeking full-time employment as economists.

hit the women Ph.D.s very hard. Over 20 percent of the women with new Ph.D.s found employment outside the United States.

As last year, fewer women with new Ph.D.s than men found employment in the federal government. In other respects, the employment patterns of men and women with new Ph.D.s were roughly similar. Five percent of the men and 6 percent of the women with new Ph.D.s were unemployed or out of the labor force. The differences between men and women in the proportions unemployed and in the proportions not in the labor force were small.

Considering all women economists employed in academic departments of eco-

TABLE 4—NUMBER OF FACULTY BY RANK AND TYPE OF DEPARTMENT, 1975–76, BY SEX
(Departments reporting by December 10, 1975, on 1975–76 Universal Academic Questionnaire)

| Type of Appointment, Rank, and Sex | All Departments | Highest Degree Offered: Ph.D. Chairman's Group | Ph.D. Other Departments | M.A. | B.A. |
|---|---|---|---|---|---|
| Number of departments reporting | 369 | 38 | 27 | 38 | 266 |
| Full-time faculty, tenure-track: | | | | | |
| All ranks, male and female, number | 3,320 | 960 | 661 | 437 | 1,262 |
| Professors | 1,166 | 507 | 229 | 140 | 290 |
| Associate professors | 855 | 186 | 201 | 120 | 348 |
| Assistant professors | 922 | 212 | 149 | 131 | 430 |
| Instructors | 209 | 25 | 37 | 19 | 128 |
| Other faculty ranks | 44 | 15 | 8 | 7 | 14 |
| Other | 124 | 15 | 37 | 20 | 52 |
| Female, percent of total | 6.1 | 4.7 | 5.1 | 6.6 | 7.4 |
| Professors | 2.9 | 1.1 | 3.4 | 4.2 | 4.8 |
| Associate professors | 5.5 | 2.6 | 5.5 | 8.3 | 6.0 |
| Assistant professors | 8.6 | 10.2 | 6.7 | 5.3 | 8.8 |
| Instructors | 14.4 | 21.9 | 5.4 | 26.3 | 12.5 |
| Other faculty ranks | 11.4 | 16.6 | [a] | [a] | 0 |
| Other | 4.8 | 0 | 5.4 | 0 | 7.7 |
| Full-time faculty, nontenure track: | | | | | |
| All ranks, male and female, number | 185 | 33 | 63 | 18 | 71 |
| Professors | 5 | 0 | 0 | 3 | 2 |
| Associate professors | 16 | 3 | 4 | 1 | 8 |
| Assistant professors | 39 | 7 | 2 | 9 | 21 |
| Instructors | 43 | 10 | 10 | 2 | 21 |
| Other faculty ranks | 10 | 2 | 3 | 3 | 2 |
| Other | 72 | 11 | 44 | 0 | 17 |
| Female, percent of total | 8.1 | 12.1 | 1.6 | 22.2 | 7.0 |
| Professors | [a] | 0 | 0 | [a] | 0 |
| Associate professors | 6.2 | [a] | 0 | 0 | 0 |
| Assistant professors | 10.2 | [a] | 0 | [a] | 14.3 |
| Instructors | 9.3 | 10.0 | 10.0 | 0 | 9.5 |
| Other faculty ranks | 10.0 | [a] | 0 | 0 | 0 |
| Other | 0 | 0 | 0 | 0 | 0 |
| Part-time faculty: | | | | | |
| All ranks, male and female, number | 564 | 106 | 111 | 63 | 284 |
| Professors | 44 | 10 | 1 | 8 | 25 |
| Associate professors | 42 | 2 | 3 | 7 | 30 |
| Assistant professors | 84 | 3 | 9 | 13 | 59 |
| Instructors | 181 | 13 | 40 | 30 | 98 |
| Other faculty ranks | 167 | 73 | 42 | 5 | 47 |
| Other | 46 | 5 | 16 | 0 | 25 |
| Female, percent of total | 14.5 | 6.6 | 18.9 | 14.3 | 14.4 |
| Professors | 4.5 | 10.0 | 0 | 0 | 4.0 |
| Associate professors | 11.9 | 0 | 0 | [a] | 13.3 |
| Assistant professors | 11.9 | [a] | [a] | 15.3 | 6.7 |
| Instructors | 14.4 | 7.6 | 15.0 | 16.7 | 14.3 |
| Other faculty ranks | 1.9 | 4.1 | 14.3 | [a] | 23.4 |
| Other | 39.1 | [a] | 37.5 | — | 28.0 |

[a] Percentage not shown when fewer than 10 in cell.

TABLE 5—DISTRIBUTION OF DEPARTMENTS OF ECONOMICS BY NUMBER OF WOMEN FACULTY MEMBERS, 1975–76

(Departments reporting by December 10, 1975 on 1975–76 Universal Academic Questionnaire)

| Number of Women on Faculty | All Departments | Chairman's Group |
|---|---|---|
| Number of departments reporting one or more full-time faculty | 378 | 38 |
| Percent distribution by number of women full-time faculty | 100.0 | 100.0 |
| All male | 63.0 | 23.7 |
| 1 woman | 27.4 | 55.3 |
| 2 women | 6.1 | 13.1 |
| 3 or more women | 3.5 | 7.9 |
| Total number departments | 392 | 38 |
| Percent distribution of departments by number women part-time faculty | 100.0 | 100.0 |
| No part-time faculty | 50.5 | 36.8 |
| 1 or more part-time faculty | 49.5 | 63.2 |
| All male | 35.8 | 44.8 |
| 1 woman | 11.7 | 18.4 |
| 2 women | 1.0 | 0 |
| 3 or more women | 1.0 | 0 |
| Total number of faculty | 3,963 | 1,132 |
| Full-time | 3,409 | 1,028 |
| Percent women | 5.9 | 4.4 |
| Part-time | 554 | 104 |
| Percent women | 12.5 | 6.7 |

nomics, women comprised only 6 percent of all full-time faculty appointments in tenure-track positions—3 percent of the full professors, nearly 6 percent of the associate professors, 9 percent of the assistant professors, and 14 percent of the instructors (Table 4). This is an increase over last year's report, particularly at the instructor and assistant professor level. Departments in the Chairman's Group have improved their proportion of women at the assistant professor level (10 percent this year compared with 7 percent last year) and at the associate professor level (nearly 3 percent this year compared with 1 percent last year). Although the differences are small, the direction of the movement is important.

Overall, considering all levels of faculty appointments, the 6 percent of the full-time faculty positions held by women (Table 5) was the same as last year. The proportion of women among those holding part-time faculty appointments fell to 12 percent this year. (Last year it was 18 percent.) The number of economics departments with all male full-time faculty has increased. In 1975–76, the proportion was 63 percent. Last year 59 percent of the departments had only male full-time faculty. This is an unfortunate regression. On the other hand, among departments in the Chairman's Group, the proportion with all male full-time faculty dropped to 24 percent. Last year it was 30 percent.

Among the 378 departments reporting in the 1975–76 survey, there were 3,409 full-time faculty positions and 554 part-time faculty positions as economists. The number of new faculty hired in 1975–76 exceeded the number of faculty released at the end of 1974–75 by only 36 full-time positions and 85 part-time (Table 6). This represented only a 1 percent increase in full-time positions and an 18 percent increase in part-time positions. In last year's survey only a 2 percent increase in full-time positions was reported and a 19 percent increase in part-time positions. Of the 38 net increase in the number of full-time positions this year, 8 were women. Of the new hires, 9 percent were women. Of those released, 8 percent were women. This is a tight market. Women at least held their small relative position. In last year's survey the net increase in the number of full-time faculty positions was 86, and 13 were women.

Among departments in the Chairman's Group, the net increase in the number of women with full-time faculty positions increased by 5 at the same time the total number of positions decreased by 4. The increases for women were at the lower faculty ranks. For men, the changes were

TABLE 6—NET CHANGE IN FACULTY POSITIONS FROM END OF 1974–75 TO 1975–76, BY SEX, ALL DEPARTMENTS AND CHAIRMAN'S GROUP

(Departments reporting by December 10, 1975 on 1975–76 Universal Academic Questionnaire)

| Item | All Ranks | Professors | Associate Professors | Assistant Professors | Instructors | Other Faculty Rank | Others |
|---|---|---|---|---|---|---|---|
| All departments: | | | | | | | |
| Faculty released end of AY 1974–75:[a] | | | | | | | |
| Full-time, number | 305 | 61 | 37 | 143 | 46 | 11 | 7 |
| Women as percent of total | 7.9 | 1.6 | 8.8 | 5.6 | 21.7 | 9.1 | b |
| Part-time, number | 95 | 3 | 5 | 12 | 42 | 32 | 1 |
| Women as percent of total | 24.2 | 0 | b | 25.0 | 23.8 | 25.0 | 0 |
| New hires, faculty, AY 1975–76: | | | | | | | |
| Full-time, number | 341 | 16 | 30 | 177 | 89 | 19 | 10 |
| Women as percent of total | 9.4 | 0 | 3.3 | 10.7 | 10.1 | 15.8 | 0 |
| Part-time, number | 180 | 2 | 4 | 38 | 69 | 45 | 22 |
| Women as percent of total | 20.0 | 0 | 0 | 36.8 | 14.5 | 22.2 | 10.0 |
| Net change, 1974–75 and 1975–76: | | | | | | | |
| Full-time, number | +36 | −45 | −7 | +34 | +43 | +8 | +3 |
| Women, number | + 8 | − 1 | −2 | +11 | − 1 | + 2 | −1 |
| Part-time, number | +85 | − 1 | −1 | +26 | +27 | +13 | +21 |
| Women, number | +13 | 0 | −2 | +11 | 0 | + 2 | +2 |
| Chairman's group: | | | | | | | |
| Faculty released end of AY 1974–75:[a] | | | | | | | |
| Full-time, number | 92 | 25 | 16 | 34 | 8 | 3 | 6 |
| Women as percent of total | 4.3 | 0 | 0 | 5.9 | b | 0 | b |
| Part-time, number | 20 | 1 | 0 | 1 | 2 | 16 | 0 |
| Women as percent of total | 20.0 | 0 | — | 0 | b | 18.7 | — |
| New hires, faculty, AY 1975–76: | | | | | | | |
| Full-time, number | 88 | 5 | 4 | 48 | 14 | 9 | 8 |
| Women as percent of total | 10.2 | 0 | 0 | 10.4 | 14.3 | b | 0 |
| Part-time, number | 31 | 0 | 0 | 4 | 7 | 13 | 7 |
| Women as percent of total | 3.2 | — | — | 0 | b | 0 | 0 |
| Net change, 1974–75 and 1975–76: | | | | | | | |
| Full-time, number | − 4 | −20 | −12 | +14 | +6 | +6 | +2 |
| Women, number | + 5 | 0 | 0 | + 3 | +1 | +2 | −1 |
| Part-time, number | +11 | − 1 | 0 | + 3 | +5 | −3 | +7 |
| Women, number | − 3 | 0 | 0 | 0 | 0 | −3 | 0 |

[a] Resignation, retirement, and nonrenewal of contracts.
[b] Percentage not shown when fewer than 10 in cell.

net decreases at the professor and associate professor level, and net increases at the lower faculty ranks. The number of women with part-time faculty positions decreased by 3 at the same time the total number of part-time faculty positions in departments in the Chairman's Group increased by 11. For both men and women these changes were primarily at the lower ranks.

Considering all the departments of economics, the women faculty employed as new hires this year tended to come less from business than did men (Table 7). This was also true last year. However, this year there was a small reversal in the previous pattern in that a slightly higher proportion of women than men came from other faculties and a slightly lower proportion of women than men came from graduate school. Among departments in the Chairman's Group, more of the women among the new hires came from government and business this year than did men,

TABLE 7—PRIOR ACTIVITY OF NEW 1975–76 APPOINTMENTS AND PRESENT ACTIVITY OF "RELEASES" FOR 1974–75, ALL DEPARTMENTS AND CHAIRMAN'S GROUP, BY SEX

(Departments reporting by December 10, 1975 on 1975–76 Universal Academic Questionnaire)

| Highest Degree Offered by Department and Activity of Faculty | New Hires in 1975–76[a] by Prior Year Activity | | Those Released for 1975–76 by Present Activity[a] | |
|---|---|---|---|---|
| | Male Percent | Female Percent | Male Percent | Female Percent |
| All departments | 100.0 | 100.0 | 100.0 | 100.0 |
| Faculty at another school in U.S. | 22.5 | 24.0 | 39.5 | 32.7 |
| Graduate student | 57.3 | 54.0 | 8.8 | 12.2 |
| Post-doctorate | 2.2 | 2.0 | 0 | 0 |
| Research institution in U.S. | 2.5 | 2.0 | 2.5 | 2.0 |
| Bank or finance in U.S. | 0.8 | 0 | 1.3 | 2.0 |
| Business and industry in U.S. | 7.2 | 2.0 | 10.1 | 10.2 |
| Federal/state government in U.S. | 5.0 | 6.0 | 14.3 | 14.3 |
| Outside United States | 1.7 | 0 | 6.3 | 8.2 |
| Retired | 0 | 0 | 10.5 | 6.1 |
| Unemployed | 0.8 | 6.0 | 0.8 | 8.2 |
| Unknown | 0 | 4.0 | 5.9 | 4.1 |
| Chairman's Group | 100.0 | 100.0 | 100.0 | 100.0 |
| Faculty at another school in U.S. | 12.6 | 9.1 | 44.1 | 16.7 |
| Graduate student | 81.0 | 63.6 | 13.0 | 25.0 |
| Post-doctorate | 4.2 | 0 | 0 | 0 |
| Research institution in U.S. | 0 | 0 | 0 | 8.3 |
| Bank or finance in U.S. | 0 | 0 | 1.3 | 0 |
| Business and industry in U.S. | 0 | 9.1 | 15.6 | 16.7 |
| Federal/state government in U.S. | 1.1 | 18.2 | 10.4 | 25.0 |
| Outside United States | 1.1 | 0 | 0 | 0 |
| Retired | 0 | 0 | 10.4 | 8.3 |
| Unemployed | 0 | 0 | 1.3 | 0 |
| Unknown | 0 | 0 | 3.9 | 0 |

[a] Includes full-time and part-time faculty.

and a lower proportion of the women than of men were selected from graduate students.

Of the women faculty released, 8 percent were unemployed compared with only 1 percent of the men. Fewer of the women released found employment on the faculty at another university than did men.

This year as last, the persons reporting for the economics departments were asked to rank women full-time faculty by whether their salaries were above or below the departmental median for the particular rank and whether their length of service in that rank was above or below the median time at that rank for departmental faculty. Such estimates ignore how much the woman's salary is above or below the median. From other evidence we know that with increases in experience, women's salaries tend to lag behind men's. For all departments, 31 percent of the women had salaries more than $250 below the medians for their ranks (Table 8). When time in rank is considered, 8 percent of the women had salaries more than $250 below the median even though their time in rank was at or above the median length of experience for that rank in the department. It must be remembered that

TABLE 8—RELATIVE SALARIES FOR RANK AND TIME IN RANK OF FEMALE FULL-TIME ECONOMISTS, 1975–76, BY TYPE OF DEPARTMENT

(Departments reporting by December 10, 1975 on 1975–76 Universal Academic Questionnaire)

| | All Women | | Time in Rank | | |
|---|---|---|---|---|---|
| Highest Degree Offered by Department and Relative Salary for Rank | Number | Percent | Above Median, Percent | At Median, Percent | Below Median, Percent |
| All departments | 222 | 100.0 | 29.7 | 36.9 | 33.3 |
| Salary above median | 69 | 31.1 | 21.1 | 5.0 | 5.0 |
| Within $250 of median | 85 | 38.3 | 4.1 | 27.9 | 6.3 |
| Salary below median | 68 | 30.6 | 4.5 | 4.0 | 22.1 |
| Ph.D., Chairman's Group | 50 | 100.0 | 24.0 | 26.0 | 50.0 |
| Salary above median | 14 | 28.0 | 16.0 | 2.0 | 10.0 |
| Within $250 of median | 19 | 38.0 | 4.0 | 24.0 | 10.0 |
| Salary below median | 17 | 34.0 | 4.0 | 0 | 30.0 |
| Ph.D., other departments | 40 | 100.0 | 22.5 | 17.5 | 60.0 |
| Salary above median | 5 | 12.5 | 5.0 | 0 | 7.5 |
| Within $250 of median | 11 | 27.5 | 5.0 | 12.5 | 10.0 |
| Salary below median | 24 | 60.0 | 12.5 | 5.0 | 42.5 |
| M.A. | 29 | 100.0 | 17.2 | 58.7 | 24.1 |
| Salary above median | 7 | 24.1 | 17.2 | 6.9 | 0 |
| Within $250 of median | 12 | 41.4 | 0 | 38.0 | 3.4 |
| Salary below median | 10 | 34.5 | 0 | 13.8 | 20.7 |
| B.A. | 101 | 100.0 | 37.6 | 44.6 | 17.8 |
| Salary above median | 41 | 40.6 | 29.6 | 8.0 | 3.0 |
| Within $250 of median | 43 | 42.6 | 5.0 | 33.6 | 4.0 |
| Salary below median | 17 | 16.8 | 3.0 | 3.0 | 10.8 |

more than two-thirds of the women faculty members in economics covered in the 1975–76 survey reported here are at the assistant professor or lower ranks. In general, entrance level faculty positions in universities have little or no difference between men and women in salary.

Women received 9 percent of the promotions for 1975–76 (19 of the 214), whereas as noted above they comprised 6 percent of the faculty (Table 9). Of the 19 promotions for women, 5 were to full professor, 7 were to associate professor, and 7 were to assistant professor. None of the promotions of women to full professor included awarding of tenure. This may well be because the women already had tenure as associate professors. Six women were awarded tenure at the associate professor level (compared with 7 promotions to this rank). Only 1 promotion and 1 tenure award to women in the professor and associate professor level were among departments in the Chairman's Group.

Many factors affect the workings of the academic labor market and positive improvements in the opportunities opened to women economists. Critical to achieving improvements are actions by men of good will and sensitivity who help change traditionally narrow views of women's role potential and help open opportunities so women can have better educational and employment opportunities. Many of the improvements needed to combat role prejudice and sex discrimination in universities involve opening the informal network to women colleagues and greater investment in on-job training opportunities

TABLE 9—PROMOTIONS AND TENURE DECISIONS FROM 1974–75 TO 1975–76, BY TYPE OF DEPARTMENT AND SEX

(Departments reporting by December 10, 1975 on 1975–76 Universal Academic Questionnaire)

| | Persons Promoted to Rank | | Persons Given Tenure at Rank | |
|---|---|---|---|---|
| Highest Degree Offered by Department and Rank | Total Number | Female as Percent of Total | Total Number | Female as Percent of Total |
| All departments: | | | | |
| Professor | 73 | 6.8 | 17 | 0 |
| Associate professor | 111 | 6.3 | 73 | 8.2 |
| Assistant professor | 26 | 26.9 | 33 | 6.1 |
| Other ranks | 3 | 0 | 2 | 0 |
| Ph.D., Chairman's Group: | | | | |
| Professor | 25 | 0 | 6 | 0 |
| Associate professor | 22 | 4.5 | 14 | 7.1 |
| Assistant professor | 3 | [a] | 1 | 0 |
| Other ranks | 2 | 0 | 0 | — |
| Ph.D., other departments: | | | | |
| Professor | 13 | 7.7 | 2 | 0 |
| Associate professor | 26 | 3.8 | 15 | 13.3 |
| Assistant professor | 3 | [a] | 3 | [a] |
| Other ranks | 0 | — | 0 | — |
| M.A.: | | | | |
| Professor | 15 | 13.3 | 3 | 0 |
| Associate professor | 16 | 12.5 | 6 | 0 |
| Assistant professor | 4 | 0 | 4 | [a] |
| Other ranks | 1 | 0 | 0 | — |
| B.A.: | | | | |
| Professor | 20 | 40.0 | 6 | 0 |
| Associate professor | 47 | 6.4 | 38 | 7.9 |
| Assistant professor | 16 | 8.5 | 25 | 0 |
| Other ranks | 0 | — | 2 | 0 |

[a] Percentage not shown when fewer than 10 in cell.

for women. *CSWEP* will continue to work on behalf of the total Association to improve the operation of the total market for economists, including the academic market, and to improve the quality of education for women students in economics. Central to the latter goal is an increase in opportunities for women faculty at higher levels in the university so as to better utilize the capabilities of women educators.

BARBARA B. REAGAN, *Chairperson*

# REPORT OF REPRESENTATIVE TO THE INTERNATIONAL ECONOMIC ASSOCIATION

During the calendar year 1975, the International Economic Association held two conferences and one symposium.

The research conference in S'Agaro, Spain, April 21–26, 1975 was on "The Microeconomic Foundations of Macroeconomic Theory." It was prepared by a Program Committee chaired by Geoffrey C. Harcourt (Australia), with Sir John Hicks (United Kingdom), T. Negishi (Japan), Luigi Spaventa (Italy), Erich Streissler (Austria), and James Tobin (United States) as members. Three Americans presented papers: Robert E. Hall, Tjalling C. Koopmans, and Axel Leijonhufvud; another three acted as discussants: Paul Davidson, Martin Shubik, and Joseph Stiglitz.

The symposium, held in Kiel (Federal Republic of Germany), July 7–12, 1975, was on "The Organization and Retrieval of Economic Knowledge." It was prepared by Herbert Giersch (F. R. Germany) and Mark Perlman (United States) as co-chairmen of a Program Committee which included also Mark Blaug (United Kingdom), J. Meyriat (France), and Vladimir A. Vinogradov (*USSR*). Papers were given by economists as well as librarians and information-systems analysts. Participants from the United States included, besides Mark Perlman, Martin Bronfenbrenner, Ronald Coase, Otto Eckstein, Glyn Evans, Michael Gort, Harry G. Johnson, Charles Kindleberger, James Morgan, Naomi Perlman, Anna Schwartz, Carl Shoup, Joseph Spengler, and William G. Tyler.

A conference at Saltsjőbaden (Sweden), August 28–September 3, 1975, was held on "Inflation Theory and Anti-Inflation Policy." The Program Committee was chaired by Erik Lundberg and included Bela Csikos-Nagy (Hungary), Giorgio Fua (Italy), Robert J. Gordon (United States), Richard Lipsey (Canada), Michael Parkin (United Kingdom), Don Patinkin (Israel), and P. de Wolff (Netherlands). Economists from the United States, delivering or discussing papers, included William Branson, Martin Bronfenbrenner, James Duesenberry, Stanley Fischer, Marcus Fleming, Robert J. Gordon, Dwight Jaffee, and Fritz Machlup.

For the calendar year 1976 three research conferences are being prepared. The first, on "Economic Relationships between Different Economic Systems (East and West)," will be held in Dresden (German Democratic Republic), June 28–July 3, 1976. The Program Committee is chaired by Oleg Bogomolov (*USSR*) and includes Abram Bergson (United States), Michael Kaser (United Kingdom), Gunther Kohlmey (German Democratic Republic), Franz Nemschak (Austria), and Josef Pajestka (Poland) as members.

The second conference will be held in Urbino (Italy) from September 1 to 8, 1976 on "Econometric Contributions to Public Policy." The Program Committee, under the chairmanship of Richard Stone (United Kingdom), includes Emilio Fontela (Spain), Lev V. Kantorovitch (*USSR*), Janos Kornai (Hungary), Jauko J. Paunio (Finland), and Richard E. Quandt (United States).

The third conference is scheduled to be held in Teheran (Iran) September 18 to 23 on "Economic Choice of Technologies for Developing Countries." A Program Committee is being formed with K. N. Raj (India) as Chairman.

In 1977 the Fifth World Conference

will be held in Tokyo (Japan) from August 29–September 4, 1977. The topic will be "Economic Growth and Resources." The Program Committee is chaired by Edmond Malinvaud (France).

Two volumes of proceedings were published during 1975 (although the imprint states 1976 as the year of publication): *Methods of Long-Term Planning and Forecasting*, edited by Tigran S. Khachaturov; and *Economic Factors in Population Growth*, edited by Ansley J. Coale. Two other volumes are scheduled for publication in the first half of 1976: *The Economics of Public Services*, edited by Martin S. Feldstein, and *Economic Integration: Worldwide, Regional, Sectoral*, edited by Fritz Machlup. All these volumes are published in London by Macmillan and in New York by The Halsted Press.

FRITZ MACHLUP, *Chairperson*

# REPORT OF COMMITTEE ON ECONOMIC EDUCATION

In 1975, this Committee continued the implementation of a broad program laid out in its 1972 agenda for improving the teaching of economics (published in the *American Economic Review* in May, 1973). This year, special emphasis has been placed on the following:

1. The project outlined in the 1972 agenda to help university economics departments provide more effective training for Ph.D.s has been further extended. It is now in successful use at seven universities training a large number of potential teachers. Information on these experiences, on teaching materials available, and on teaching techniques has been organized into a special set of materials which the committee hopes will be useful to a wide variety of universities training future teachers. Further information can be obtained by writing Dr. Arthur Welsh, Joint Council on Economic Education, 1212 Avenue of the Americas, New York, New York 10036.

2. *Recent Advances In Economics,* edited by Rendigs Fels and John Siegfried of Vanderbilt University, under the general policy guidance of this Committee, has appeared as the first of the new *Readings* series designed to help teachers at small institutions keep up-to-date on central areas of economic theory and policy. This volume contains 25 articles chosen to provide information on recent advances in economics. They are selected to be of special help to teachers of economics at community and junior colleges and at other smaller colleges and universities where faculty members often find it difficult to keep up with new theory and policy developments and do an adequate job teaching a wide variety of courses. The volume is available (in paperback) from Richard D. Irwin, Inc., which has published it in cooperation with this Committee. Future volumes will be prepared and published if this one seems to serve a useful purpose.

3. The Committee has completed a review of experience with the *Journal of Economic Education,* published by the Joint Council on Economic Education with the cooperation of this Committee. The Journal has, in our judgment, been very successful. It now has a paid subscription list of some 2,600 persons, mainly college and university teachers at the undergraduate level. It is now in its seventh year, and has published a wide-ranging set of articles, mainly reporting research results evaluating different teaching approaches for undergraduate courses in economics. It has become the main source of information to members of the profession on such teaching developments and experiments. Professor Henry Villard, editor of the *Journal,* plans to broaden out somewhat the types of articles published, including a number of commissioned pieces on major developments of special interest to undergraduate teachers. We commend the *Journal* to economists who have special interests in undergraduate teaching.

4. The Joint Council on Economic Education is publishing reports on a series of six special projects, at as many universities, to develop experimental courses in elementary economics which might be transferable to smaller colleges which, because of limited manpower or time, have difficulty in developing elementary courses of particular usefulness to their own students. A small volume is being published on the course developed at each of the

cooperating institutions (Colorado, Florida State, Harvard, Indiana, University of Nebraska, and Vanderbilt). The courses developed at the different institutions vary widely. Copies of the booklets are available from the Joint Council on Economic Education, 1212 Avenue of the Americas, New York, New York.

5. This Committee has cooperated with *Change* Magazine, under a grant from the Fund for Post-Secondary Education, in a project to publish detailed information on interesting examples of innovative undergraduate teaching in a number of fields of knowledge. The purpose of the program is to identify, with the cooperation of the various disciplinary associations, examples of some of the best efforts to improve undergraduate teaching. *Change* (and the *Journal of Economic Education*) will publish the studies of each of the programs chosen for special attention. Choices of the programs to be included will be made by the officials of the project. These studies should be published over the period 1976–78, and, tentatively, will include several economics courses.

G. L. BACH, *Chairperson*

# REPORT OF THE COMMITTEE ON U.S.-SOVIET EXCHANGES

It may be desirable for the record to begin this report with an account of the early discussions leading to the visit of the Soviet delegation in April 1975.

### Preliminary Talks and Letters

The first suggestion for this undertaking came from Academician Nikolay P. Fedorenko, Director of the Central Economical Mathematical Institute of the Academy of Sciences of the *USSR*. He made the suggestion at a small party he gave for Fritz Machlup and others in Moscow in December 1972 on the occasion of a conference of the International Economic Association. Machlup acknowledged Fedorenko's "suggestion of annual symposia of the *USSR* and the U.S.A." in his letter of December 22 and promised further explorations with the Ford Foundation about the possibilities of financing such a project. A week later at the Executive Committee of the American Economic Association (*AEA*), Machlup reported the discussions with Fedorenko and was asked to continue his explorations with the Ford Foundation. He and Abram Bergson met with officers of the Ford Foundation. On March 15, Machlup wrote to Fedorenko that the Executive Committee of the *AEA* and the officials of the Ford Foundation expressed considerable interest in the project. Fedorenko replied on April 10 and promised to make concrete proposals after consulting his colleagues of the Academy of Sciences of the *USSR*. Machlup wrote to Fedorenko on May 8, 1973, but received no reply.

In August 1973, on the occasion of another visit of Machlup in Moscow, Academician Tigran S. Khachaturov repeated the proposal for annual symposia (apparently without earlier discussion with Fedorenko). He suggested that an official invitation be sent by the President of the American Economic Association, addressed to President Mystislav Keldish of the Moscow Academy, with carbon copies to Fedorenko and Khachaturov. Machlup told Khachaturov that we would first have to secure the funds and then obtain the definite concurrence of the Executive Committee of the *AEA*.

In October 1973, Machlup wrote to Craufurd Goodwin of the Ford Foundation, requesting some assurance of Foundation help before the proposal could be brought to the Executive Committee. On December 12, 1973, in a meeting at the Ford Foundation, it was agreed that *IREX* would provide the funds needed by the *AEA* to host twelve economists from the *USSR* for two weeks in the United States. At the December meetings Machlup reported again to the Executive Committee of the *AEA*, which responded favorably, and in a letter of January 6, 1974, he briefed President Walter Heller on the course of events.

In February 1974, President Heller applied to Allen Kassof of *IREX* for a grant of $7,000 to finance the proposed visit. After a concurring vote by the Executive Committee in its March meeting, Heller appointed a committee to plan for the visit of the *USSR* economists, the committee to consist of Abram Bergson, John R. Meyer, and Fritz Machlup (as Chairman). On May 10, Heller dispatched an invitation to President Keldish of the *USSR* Academy, with carbon copies to Fedorenko and Khachaturov, and (through Kassof of *IREX*) to Fedoseëv in Moscow. He also asked Burkhardt of

the American Council of Learned Societies (*ACLS*) in Washington to endorse the invitation in accordance with the agreement with the *ACLS* under the provisions of the General Agreement between the two governments concerning contacts, exchanges, and cooperation, signed on June 13, 1973.

No official reply was received for several months. In August Machlup saw Fedorenko and Khachaturov at the Budapest Congress of the International Economic Association. Khachaturov asked whether February or March 1975 would be acceptable dates for a visit of a Russian delegation. On October 20, 1974, Khachaturov sent to Heller an official acceptance of the invitation. This was confirmed by Heller in his letter of November 8.

At the Executive Committee in December 1974 it was agreed to propose scheduling the Symposium for April 1975. Machlup cabled to Khachaturov and received confirmation on January 6, 1975. At last the preliminaries had been completed and the actual planning for the first Symposium could get under way.

### Planning the First Symposium

In a letter of January 20, 1975, Khachaturov proposed six topics. All of these topics were accepted. Khachaturov was asked to send a list of the members of the Russian participants and to send copies of the papers as soon as possible. A list of twelve economists was given by Khachaturov in a letter postmarked March 14 and again in a telegram of March 25. The name of a Mr. Smirnov, included in the first list, was omitted in the telegram; upon our inquiry we were told by telegram on April 1 that Smirnov had withdrawn. The arrival was set for Friday, April 11, at Dulles Airport, Washington.

Only four papers had been received, two of them too late for distribution to prospective participants. Here is a list of the four papers:

(1) T. S. Khachaturov, "Dynamics of Capital Investments and Their Efficiency;"
(2) S. Nikitin, "Capital Investments, Fixed Productive Assets, and Their Efficiency: A Comparative Study;"
(3) A. Tolkachev, "Methods of the Long-Term Forecasting of the Volume and Effectiveness of Investments in the National Economy;" and
(4) O. Bogomolov, "Experience in Improving Economic Management in the European Socialist Countries."

Only ten of the eleven announced members of the *USSR* delegation actually came. The eleventh man, Victor Petrovich Krasovsky, Department Chief, Institute of Economics, did not arrive; he evidently failed to receive the necessary exit visa.

John Meyer appointed Gary Fromm of the National Bureau of Economic Research (*NBER*) as coordinator of the program and itinerary. The program provided for meeting the Russian visitors at their arrival at Dulles Airport, on Friday, April 11; an agenda meeting and general discussions on the morning of Saturday, April 12; a reception by the *NBER* in Washington on Sunday, April 13 in the afternoon; and six half-day sessions for papers and discussions; Sunday morning, Monday morning, Tuesday morning and afternoon, Wednesday morning and Thursday morning, all held in the State Department in Washington. For the subsequent seven days it provided for a visit to Columbia University on Friday, April 18, followed by two free days in New York City; a visit to Princeton University on Monday, April 21; to the Wharton School of the University of Pennsylvania on Tuesday, April 22; to Harvard University and the *NBER* in Cambridge on Wednesday, April 23; to Massachusetts

Institute of Technology and Data Resources Inc. on Thursday, April 24. Departure from Washington was set for Friday, April 25.

### The Symposium

The following topics were discussed in the six sessions of the symposium:

(1) "Methods of Determining the Efficiency of Investments and the Choice of Variants and Alternatives"
(2) "Methods of Forecasting Capital Investments"
(3) "Experience in Improving Economic Management in European Socialist Countries"
(4) "Effectiveness of Capital Investment in Public Service Sectors"
(5) "Comparisons of the Efficiency or Productivity of Investments in Different Countries"
(6) "Prospects for Soviet-American Economic Relations"

The composition of what might be called the American delegation proved to be a serious problem. The main difficulty was that, of the large number of persons invited from places other than Washington, only a few could find the time to attend at all and practically none could attend for more than two days. Only economists residing in Washington could be counted on to attend regularly, but a number of those approached were also unable to participate. As a result, the participation by American economists was totally inadequate, especially after the first meeting. Some of our colleagues from Philadelphia, particularly Lawrence Klein and Irving Kravis saved the day by coming to one additional session. Several economists from the Washington region did the best they could to keep up the appearance of a symposium in some of the remaining sessions.

No one is to be blamed for this poor showing on our part. We were evidently too optimistic regarding the length of time for which our colleagues could get away from their regular duties to attend a meeting of which they had received very short notice. With hindsight one may say that we should have relied more on our colleagues residing in the Washington area and should have invited them much earlier than we did.

Some of our Russian guests remarked that papers should have been prepared also by the American participants. This, however, would have been impossible because of the short notice. Even if all Russian papers had arrived in time for distribution four or five weeks before the symposium, one could not have expected any of the American economists to prepare discussion papers in a hurry and have them available for the Russian visitors.

### The Visits to East Coast Institutions

If the formal sessions of the symposium proved less than successful, the visits to various institutions can be regarded as highly satisfactory. Initially our guests were somewhat disappointed when they were told that a visit to the West Coast could not be arranged; there was neither enough time nor enough money available for such a long trip. However, they found the visits to the eastern institutions very successful from several points of view; they were glad to see some of the sights, to meet many economists, to listen to lectures and briefings, and to engage in informal discussions on topics of their interests.

In New York, the visitors spent the morning with the Economics Department at Columbia University. In the afternoon they attended a seminar at the National Bureau of Economic Research. There

were no planned activities for Saturday and Sunday, so that the visitors could make their own plans for these free days.

In Princeton, our guests attended two lectures in the Department of Economics, were shown the Computer Center and the Woodrow Wilson School of Public and International Affairs, and, finally, the Institute for Advanced Study.

In Philadelphia, they received briefings at the Wharton Econometric Forecasting Associates Inc. and attended a seminar at the Economics Department of the University of Pennsylvania.

In Cambridge, the visitors had a briefing at the National Bureau of Economic Research and a discussion comparing American and Soviet methodologies and findings. They were also shown the *NBER* Computer Research Center, especially their *SESAME, TROLL,* and *XMP* systems. In addition, there were visits at Data Resources Inc., Harvard University and Massachusetts Institute of Technology.

### Acknowledgments

The American Economic Association should express its gratitude to *IREX* for financing the entire undertaking. Although originally only $7,000 had been granted, Mr. Kassof and Mr. Matuszewski of *IREX* increased the funds at our disposal to a total of $13,000. This was highly necessary, especially as our visitors had to be wined and dined in the company of American colleagues selected to meet them at various places.

The officials and the staff of the *NBER* deserve high praise and sincere gratitude for making all the arrangements in Washington and elsewhere, including the arrangements for interpreters, guides, hotel accommodations, restaurants, travel, etc. A special vote of gratitude is owed to Gary Fromm of the Washington Office of the *NBER,* who personally attended to many of the organizational details.

### Recommendations

Khachaturov proposed that the next symposium be held in Moscow in the second half of June 1976. The timing may suit many American economists, since it will be well after commencement here and before some other conferences in Europe. (We know of at least two European conferences that begin on June 28, 1976.)

Khachaturov listed a number of preferred subjects for the symposium. Bergson suggested, and the rest of the committee agrees, that we should not embark on a wide array of topics, but choose one that in itself is broad enough to elicit interested response from a number of specialists. The topic most appropriate from this point of view would be "The Economics of Technological Progress."

This broad subject would allow discussions of such aspects as the meaning and nature of technological progress; the institutional framework, including the organization of research and development, patenting, etc.; trends in the volume of *R&D* activities; the measurement of technological progress; the diffusion of new technologies; and the international transfer of technological information.

We believe that a good group of American economists, knowledgeable in several aspects of this topic, could be selected and invited to the symposium. As to the number of participants, we think of ten to twelve. A large number would not only make travel expenses too high but might reduce the intensity of involvement in the discussions.

Khachaturov wants us to submit papers in advance. We fear that such a requirement would reduce the chances of our best possible candidates accepting our in-

vitation. Hence we propose that American participants be asked to submit only brief outlines, which they would then elaborate orally at the symposium in Moscow.

Fromm suggested that we try to obtain funding of the travel expenses from the U.S. Department of State. We propose that this be explored. If it proves impossible, we can still fall back on *IREX* as the financial sponsor of the undertaking.

We finally propose that a committee be appointed without much delay, charged with the task of selecting the American participants and assigning the topics on which they ought to prepare outlines for the symposium. We ought to inform Khachaturov before the summer is over.

FRITZ MACHLUP, *Chairperson*

# The London *Economist* urges this book for the "earnest citizen." Why not for your good and earnest students?

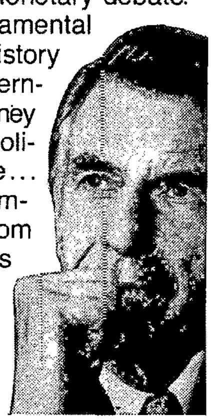
# Money

## WHENCE IT CAME, WHERE IT WENT

■ "If an earnest citizen, stirred to interest in monetarism... determines to try to get to the bottom of the subject, his instinct is absolutely sound. No area of public policy is likely to affect his future prospects more directly than the ability of the economy to cope with inflation or deflation or the modern malady of contriving to have both at once.

"Yet if the instinct is sound, the pursuit of information is perplexing in the extreme...

"But he might not... [find it] so if he could be persuaded to pick up Professor Galbraith's light-hearted yet profoundly serious gallop through the history of money which, beginning with the Kings of Lydia and ending with the slumpflation of 1975, might give him, as a general reader, some sense of the simplicities, ironies and deceptions of the world's long monetary debate. There he will learn a fundamental point — that throughout history what can be done by governments or bankers about money supply is fundamentally a political, not an economic issue... The conclusion will not commend the book to those whom Professor Galbraith defines as the conventionally wise. That is hardly his intention. But **the interested citizen will find himself enlightened and, which is even more unlikely, continuously amused."**



—*The Economist,* London, October 25, 1975

■ "MONEY is an application of Galbraith's analysis of power to the economic history of the West. Who wins and who loses is always clear in this highly readable account of the evolution of banking, the rise of paper money, the role of war in inflation, the mechanics of the great swindles from John Law to our own Bernard Cornfield, the rise of Keynesian fiscal theory, and the mystique and muddle which accompany the activities of the Federal Reserve and other central banks.

**"Galbraith performs the valuable service of draining the mystery out of banking... As ever, his prose, however gloomy its message, is a literary tonic."**

—ROBERT LEKACHMAN, *The Washington Post,* September 7, 1975

# by John Kenneth Galbraith

**$10 • HOUGHTON MIFFLIN COMPANY**

2 Park St., Boston 02107

Publisher of the American Heritage Dictionary

Photo: Murray Mosher

Begin making plans to attend the

# *Eighty-ninth*
# Annual Meeting of The American Economic Association

to be held in

# ATLANTIC CITY, NEW JERSEY

Thursday, Friday, Saturday
**SEPTEMBER 16-18, 1976**

---

The Employment Center opens Wednesday, September 15.

---

See the Notes section of the June, AER for the preliminary program.

---

The 1977 meeting will be held in New York City, December 28-30.

---

# The American Economic Review

## ARTICLES





JUNE 1976

# THE AMERICAN ECONOMIC ASSOCIATION

Founded in 1885


• Published at George Banta Co., Inc., Menasha, Wisconsin.

• THE AMERICAN ECONOMIC REVIEW, including four quarterly numbers, the *Proceedings* of the annual meetings, the *Directory*, and Supplements, is published by the American Economic Association and is sent to all members five times a year, in March, May, June, September, and December.

Association membership dues for 1976, which include a subscription to both The *American Economic Review* and the *Journal of Economic Literature*, are as follows:

$25.00 for regular members with rank of assistant professor or lower, or with annual income of $12,000 or less;

$30.00 for regular members with rank of associate professor, or with annual income of $12,000 to $20,000;

$35.00 for regular members with rank of full professor, or with annual income above $20,000;

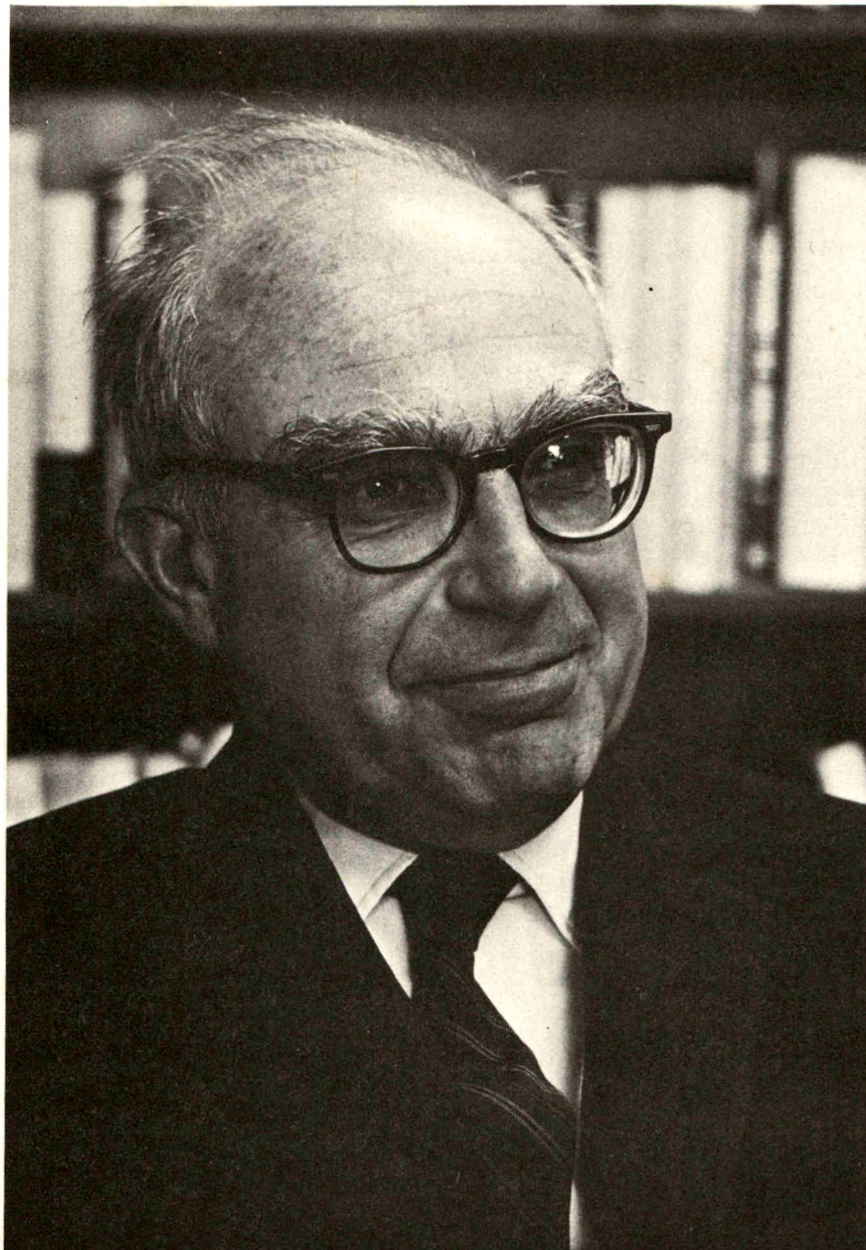
$12.50 for junior members (registered students). Certification must be submitted yearly.

Subscriptions (libraries, institutions, or firms) are $37.50 a year. Only subscriptions to both publications will be accepted. Single copies of either journal may be purchased from the Secretary's office, Nashville, Tennessee.

For foreign postage, add: PUAS countries, $3.20; all other countries and Canada, $3.75.

• Correspondence relating to the *Papers and Proceedings*, the *Directory*, advertising, permission to quote, business matters, subscriptions, membership and changes of address should be sent to the Secretary, C. Elton Hinshaw, 1313 21st Avenue South, Nashville, Tennessee 37212. Change of address must reach the Secretary at least six (6) weeks prior to the month of publication. The Association's publications are mailed second class.

• Second-class postage paid at Nashville, Tennessee and at additional mailing offices. Printed in U.S.A.



Moses Abramovitz

# MOSES ABRAMOVITZ

## Distinguished Fellow

## 1976

Moses Abramovitz is an economist for all seasons. Insight, perseverance, and modesty are a rare combination; and they have made him into a rare economist. His scholarly temperament made him shun the economic fashions; but thanks to his sense of the important, his work usually became the forerunner of fashions. While others imported analytic tools into economic history, he brought an historian's painstaking thoroughness and respect for evidence to economic analysis; and we owe to him much of what we know about our economy's growth and its ups and downs.

His dissertation was a forerunner of modern dynamic analysis. Among the first to make use of measures of total factor productivity growth, he identified the residual as a measure of our ignorance, and, as a means of reducing that ignorance, he urged recognition of health, education, training, and research as costly inputs contributing to long-term growth.

If his insight gave him the first word on much that has since become the economist's standard equipment, his perseverance also gave him the last word on many topics. He was the one definitely to establish inventory adjustment as the major component of cyclical instability; and the theory of long swings, the second major component, is also largely his work, based on his meticulous study of construction activity over a hundred years. We are pleased to recognize his distinction.

THE AMERICAN ECONOMIC REVIEW

June 1976

VOLUME 66, NUMBER 3



• Manuscripts and editorial correspondence relating to the regular quarterly issue of this REVIEW should be addressed to George H. Borts, Managing Editor of THE AMERICAN ECONOMIC REVIEW, Brown University, Providence, R.I. 02912. Manuscripts should be submitted in duplicate and in acceptable form and should be no longer than 50 pages of double-spaced typescript. A submission fee must accompany each manuscript: $15 for members; $30 for nonmembers. *Style Instructions* for guidance in preparing manuscripts will be provided upon request to the editor.



**Articles**

**Shorter Papers**

# Flexible Exchange Rates, Forward Markets, and the Level of Trade

*By* David P. Baron*

One of the arguments made against a system of flexible exchange rates for regulating the balance of payments of countries is that the uncertainty posed by exchange rate variation will reduce the level of trade. Proponents of such a system have argued that traders are able to hedge their exchange risks by using forward exchange contracts or similar covering arrangements, and thus, traders can behave as if the exchange rates were certain. While there are numerous other arguments for and against a system of flexible exchange rates, the direct effect of exchange rate uncertainty on trade is central to the choice between the systems. The purpose of this paper is to investigate in the context of a simple model the effect of uncertain exchange rates on the output and exports of firms, and to determine the role of forward markets in planning output and trade. The model is based upon recent work concerning incomplete capital markets by Peter Diamond, Steinar Ekern and Robert Wilson, Ekern (1974, 1975), Hayne E. Leland (1973, 1974), Roy Radner, and Joseph Stiglitz.

It will be shown that in equilibrium all investors are indifferent to the covering decisions of firms, and hence, that the values of all firms are independent of the covering of any firm. In addition, all firms are able to plan their production and exports on the basis of the forward exchange rates and will behave as a multiproduct, "profit" maximizing firm using the fully covered marginal profit as a "planning-equivalent" even though the firm may not fully cover its exchange risks. The level of trade with a flexible exchange rate system will thus be identical to that which would exist with fixed exchange rates equal to the forward rates, so this uncertainty regarding exchange rates has no effect on trade.

To be able to plan its output levels, a firm does not need any information regarding the preferences or probability assessments of its shareholders, yet all shareholders will unanimously prefer the production levels that maximize the fully covered profit. A voting process may be constructed that will lead to the equilibrium outputs, but such a process is not necessary, since a firm can investigate the preferences for various output levels by considering a hypothetical shareholder. The analysis for the hypothetical shareholder indicates unanimity on the part of all actual shareholders thus permitting the planning of outputs and exports.

The separation of control and ownership of firms creates the possibility of alternative criteria for choices of output levels with those criteria reflecting the preferences of managers or the managers' perception of what would be in the best interests of its shareholders. Managers may believe, for example, that they have more accurate expectations regarding future exchange rates than have shareholders, and the managers may then wish to use their

* Professor of managerial economics and decision sciences, Graduate School of Management, Northwestern University. This research was undertaken while I was professeur associé, Institut d'Administration des Entreprises, Université de Droit, d'Economie et des Sciences, Aix-en-Provence, France.

probability assessments and maximize, for example, the expected utility of profits for some utility function and their expectations. The output levels resulting with this criterion will be shown to be identical to those resulting from maximizing fully covered profit, and thus managers of firms who use their own expectations and preferences for profit will be acting in the best interests of the shareholders of the firm. The separation of ownership and control thus does not affect trade if the managers maximize the expected utility of profit and are able to cover their exchange risks.

Central to these results is the existence of a forward exchange market or its equivalent for the currency of each country to which the firm exports. If investors are able to borrow and lend foreign currencies, the existence of forward markets is not necessary, since the interest rate-parity theorem of arbitrage is essentially a statement of the redundancy of forward exchange markets when foreign currency borrowing and lending is possible. If forward exchange markets or their equivalents do not exist, the indifference and unanimity results may not hold.

The model and the investor equilibrium are presented in Section I, and the production planning of firms and the resulting levels of trade are considered in Section II. The role of forward markets is investigated in Section III, and conclusions are offered in the final section.

## I. The Model and an Investor Equilibrium

### A. *The Capital and Foreign Exchange Markets*

Individual investors will be assumed to have the opportunity to trade in both foreign exchange markets and capital asset markets for the shares of firms. To simplify the notation, it will be assumed that there is only one firm in each country with its shares traded only in the capital asset market in that country. Investors will initially be permitted to purchase shares in any of the capital asset markets, all of which are assumed to be competitive in that the market participants do not believe that their portfolio decisions will affect share prices. The capital asset markets are assumed to be incomplete in the sense that the firm through its production decisions may offer to investors patterns of returns that are not found elsewhere in the markets and cannot be created independently by investors. Transactions costs are taken to be zero and short sales are permitted.

In addition to investing in shares, individuals may trade in the foreign exchange markets as may firms that wish to hedge their accounts receivable in the forward exchange market. A two-period model will be utilized, so three exchange rates are involved. The period-1 spot exchange rate between country $n$ and $j$ is denoted by $r_{nj}$ and is measured in units of currency $n$ per unit of currency $j$. Firms and investors may also sign contracts to exchange one currency for another in period 2 at the forward exchange rate $\rho_{nj}$.[1] The future spot exchange rate in period 2 is uncertain and is denoted by $\rho_{nj}(\theta)$, where $\theta$ is a state of nature reflecting the actions of other traders, such as tourists and governments, long-term capital movements, and exogenous events. The model analyzed here does not provide an explanation of the determination of the future spot rates (see H. G. Grubel), since not all determinants of the balance of payments are considered here. The model will, however, indicate the relationship between spot and forward rates.

Forward exchange markets are assumed to exist for every pair of currencies, and the exchange rates for a country with itself equals one ($\rho_{nn}(\theta)=\rho_{nn}=r_{nn}=1$ for all $\theta$

[1] Other forms of forward exchange transactions such as swaps will not be considered, since they simply duplicate the forward exchange markets.

and $n$). Two assumptions regarding the foreign exchange markets will be made. First, arbitrage opportunities within a market yield a zero return, so that $r_{nj}=1/r_{jn}$ and $r_{nj}=r_{nl}r_{lj}$ for all $n$, $j$, and $l$, and the same is true for the forward rates and the future spot rates in every state $\theta$. Investors are able to borrow and lend in foreign currencies, so arbitrage between spot and forward markets is permitted. Second, the foreign exchange markets are assumed to be competitive in the sense that investors do not believe that their foreign currency trading or the actions of any firm will affect the exchange rates.

## B. *Firms*

In order to focus on the effects of exchange rate uncertainty, the revenue and cost functions of firms are assumed to be deterministic and known to all market participants. The firm located[2] in country $n$ is taken to sell both in its home country and in a set of foreign countries with the firm's decisions being made in period 1 and the proceeds from its production and financial decisions being realized in the second period. Let there be $n=1, \ldots, N$ countries and assume that all foreign sales are exports from the home country that are invoiced in the currency of the importing country.[3] The quantity of goods sold in country $j$ by the firm located in country $n$ is denoted by $q_{nj}$, and the price in that country is given by an inverse demand function $p_{nj}(q_{nj})$. The quantity $q_{nj}$ is assumed to be sold on a line-of-credit basis with payment to be received in period 2.[4] The total currency $j$ revenue $R_{nj}$ is thus $R_{nj}=p_{nj}(q_{nj})q_{nj}$, and that amount is exposed to uncertainty regarding the spot exchange rate that will obtain at the time at which the firm receives payment for its exports. The firm may hedge its foreign currency risk by selling currency $j$ against its home currency at the forward exchange rate $\rho_{nj}$. Letting $A_{nj}$ be the forward sale ($A_{nj}>0$) or purchase ($A_{nj}<0$) of currency $j$ against $n$, where $A_{nj}$ is measured in currency $j$, the net exposure in currency $j$ is $(R_{nj}-A_{nj})$. The total revenue $TR_n(\theta)$ collected in period 2 in the firm's home currency is the net exposure converted at the future spot exchange rate plus the proceeds from the forward transactions for each country or

$$(1)\quad TR_n(\theta) = \sum_j \left[\rho_{nj}(\theta)(R_{nj} - A_{nj}) + \rho_{nj}A_{nj}\right]$$

where $A_{nn}\equiv 0$.[5] The total home currency revenue is thus the uncovered revenue plus the return on speculation $(\rho_{nj}-\rho_{nj}(\theta))A_{nj}$. To simplify the analysis, all production will be assumed to take place in the home country and all factor inputs are assumed to be purchased in the currency of the home country. The cost of production will be represented by a cost function $C_n(q_n)$, where $q_n=(q_{n1}, \ldots, q_{nN})$, with payment to the factors of production made in period 2.[6] The profit $\pi_n(\theta)$ realized in period 2 in state $\theta$ is thus

$$(2)\qquad \pi_n(\theta) = TR_n(\theta) - C_n(q_n)$$

[2] The term "located" is intended to indicate that the firm converts its revenue into the currency of that country and distributes its profits to its investors in that currency.

[3] The majority of exports are invoiced in the currency of the exporting currency although the practice varies considerably among countries and across industries as indicated by Sven Grassman. Instead of representing each firm as invoicing certain exports in its home currency and others in foreign currencies as well as importing factor inputs invoiced in a variety of currencies, the common assumption will be made that all exports are invoiced in the currency of the importing country.

[4] Grassman found that the average payment period for Swedish exports was 91 days in 1968.

[5] Firms will not be considered to conduct arbitrage operations or to deal in the forward markets not involving their country. Firms are also assumed not to borrow and lend in foreign currencies. The significance of this assumption will be considered in Section IV.

[6] The model may be generalized readily to the case in which a firm has production units in a foreign country. For example, if firm $n$ has a production unit in a

The profit function is assumed to be strictly concave and continuously differentiable in every state.

### C. *Investors*

Investors are assumed to maximize subject to a budget constraint their expected utility of return from investments in the shares of firms and from their foreign currency transactions. An investor[7] in country $m$ has an initial endowment consisting of shares of the equity of firms in each country and savings (or borrowing) in each currency. The investor's initial share of the equity of the firm in country $n$ will be denoted by $\bar{z}_{mn}$, and the market value of that firm in currency $n$ will be denoted by $V_n$. The investor's initial endowment of currency $n$ will be denoted by $\bar{y}_{mn}$ which has a current value in the home currency of $r_{mn}\bar{y}_{mn}$ with $\bar{y}_{mn}>(<)0$ representing savings (borrowing). The investor may also have been endowed with a contract for a proportion $\bar{t}_{mnj}$ of the forward sales of currency $j$ against currency $n$ where $\bar{t}_{mnj}A_{nj}>(<)0$ represents a forward purchase (sale) of currency $j$ against $n$. The forward purchase commits the investor to supply $\sum_j \bar{t}_{mnj}\rho_{nj}A_{nj}$ of currency $n$ in period 2. For $j=n$, $A_{nn}\equiv 0$, so $\bar{t}_{mnn}$ is not a decision variable. No margin requirement on forward contracts is assumed to be required, but the inclusion of such a requirement is straightforward and causes no changes in the results.[8]

The initial wealth $W_m^0$ of the investor is thus composed of the current value of the share holdings plus the savings (borrowing), or

$$(3) \quad W_m^0 = \sum_n \bar{z}_{mn} r_{mn} V_n + \sum_n \bar{y}_{mn} r_{mn}$$

Given this initial wealth, the investor reallocates his portfolio by choosing new ownership proportions $z_{mn}$ and new foreign currency holdings $(t_{mnj}, y_{mn})$ subject to the budget constraint

$$(4) \quad W_m^0 = \sum_n z_{mn} r_{mn} V_n + \sum_n y_{mn} r_{mn}$$

The forward commitments do not enter the budget constraint because no margin requirement is incorporated into the model.

The total return on the investor's portfolio is his share of the profits of firms, his lending and borrowing return, and his gain or loss on forward transactions all converted to the home currency in period 2. His share of the profit of a firm is converted to the home currency at the future spot exchange rate for a value of $\sum_n z_{mn}\rho_{mn}(\theta)\pi_n(\theta)$. The value of the savings (borrowing) in currency $n$ is $y_{mn}i_n$, where $i_n$ is the gross borrowing-lending rate in currency $n$, and this will be converted at the future spot rate for a total value of $\sum_n \rho_{mn}(\theta)y_{mn}i_n$. The home currency value of the forward purchases $\sum_n t_{mnj}A_{nj}$ of currency $j$ is $\rho_{mj}(\theta)\sum_n t_{mnj}A_{nj}$, and the total value of all forward purchases is $\sum_j \rho_{mj}(\theta)\sum_n t_{mnj}A_{nj}$. The cost in currency $n$ of the forward purchases of $j$ against $n$ is $\rho_{nj}t_{mnj}A_{nj}$, and the total obligations in currency $n$ are $\sum_j \rho_{nj}t_{mnj}A_{nj}$. This amount of currency $n$ must be purchased at the future spot exchange rate $\rho_{mn}(\theta)$ for a total currency $m$ cost of $\sum_n \rho_{mn}(\theta)\sum_j \rho_{nj}t_{mnj}A_{nj}$. The net gain $S_m(\theta)$ on speculation via forward transactions is thus

$$(5) \quad S_m(\theta) = \sum_n \rho_{mn}(\theta) \sum_j (\rho_{nj}(\theta) - \rho_{nj})t_{mnj}A_{nj}$$

---

country $j$ and $C_{nj}(q_{nj})$ is the cost for that unit, then the profit from country $j$ in terms of the currency of the home country is

$$\rho_{nj}(\theta)(R_{nj} - C_{nj}(q_{nj}) - A_{nj}) + \rho_{nj}A_{nj}$$

The results presented below also hold for this case.

[7] To simplify the indexing notation, individual investors in each country are not identified.

[8] If a fraction $\beta_{mnj}$ of the forward commitment is required as margin, let $\beta_{mnj}\in[0, 1]([-1, 0])$ if $\bar{t}_{mnj}\rho_{nj}A_{nj}\geq(<)0$. Then, if the margin deposit is made in the home currency of the investor, the total margin deposit is $\sum_n r_{mn}\sum_j \beta_{mnj}\bar{t}_{mnj}\rho_{nj}A_{nj}$. If this margin deposit earns the risk-free rate of interest, the margin is simply another form of saving in the home currency, and none of the results are affected.

where $(\rho_{nj}(\theta)-\rho_{nj})$ is the gain to speculation on each unit of currency $j$ purchased forward against $n$. The total portfolio return $X_m(\theta)$ in currency $m$ is then

$$(6)\quad X_m(\theta) = \sum_n z_{mn}\rho_{mn}(\theta)\pi_n(\theta) + \sum_n \rho_{mn}(\theta)y_{mn}i_n + S_m(\theta)$$

Each investor is assumed to have preferences for period 2 return that are represented by a measurable utility function $U_m(X_m(\theta))$ that is increasing, strictly concave, and continuously differentiable in $X_m(\theta)$.[9] Each individual is assumed to assess a probability distribution on $\theta$, and expectation with respect to that assessed probability distribution will be denoted by $E_m$. Investors may have different assessed distributions of the states of nature, but all are assumed to agree that there is no default risk on forward contracts or on borrowing commitments. The investor thus seeks to maximize

$$E_mU_m(X_m(\theta))$$

with respect to the $z_{mn}$, $t_{mnj}$, and $y_{mn}$ subject to the budget constraint implied by (3) and (4). In equilibrium the decisions for all investors are restricted by the constraints

$$(7)\quad \sum_m z_{mn} = 1, \quad n = 1,\dots,N$$
$$\sum_m t_{mnj} = 1,$$
$$j = 1,\dots,N,\ n = 1,\dots,N,\ n \neq j$$

### D. *Investor Equilibrium*

To develop the portfolio optimality conditions for an investor, solve the budget constraint in (4) for the savings $y_{mm}$ in the home currency and substitute into $X_m(\theta)$ in (6) to obtain the net return

$$(8)\quad X_m(\theta) = \sum_n z_{mn}[\rho_{mn}(\theta)\pi_n(\theta) - r_{mn}i_mV_n] + \sum_n y_{mn}[i_n\rho_{mn}(\theta) - i_mr_{mn}] + S_m(\theta) + \sum_n \bar{z}_{mn}r_{mn}V_n + \sum_n \bar{y}_{mn}r_{mn}$$

The first term in brackets in (8) is the net return on the shares of the firm in country $n$ less the opportunity cost of purchasing shares in lieu of saving in the home country. The second term in brackets is the home currency value of a unit of savings in country $n$ less the opportunity cost of saving in country $n$ instead of the home country. The third term is the gain or loss due to speculation using forward transactions.

The necessary optimality conditions for the investor are[10]

$$(9)\quad \frac{\partial E_mU_m}{\partial z_{mn}} = E_m\{U_m'\cdot(\rho_{mn}(\theta)\pi_n(\theta) - r_{mn}i_nV_n)\} = 0, \quad n = 1,\dots,N$$

$$(10)\quad \frac{\partial E_mU_m}{\partial t_{mnj}} = E_m\{U_m'\cdot(\rho_{mn}(\theta)(\rho_{nj}(\theta) - \rho_{nj})A_{nj})\} = 0, \quad n = 1,\dots,N,\ j = 1,\dots,N,\ j \neq n$$

$$(11)\quad \frac{\partial E_mU_m}{\partial y_{mn}} = E_m\{U_m'\cdot(i_n\rho_{mn}(\theta) - i_mr_{mn})\} = 0, \quad n = 1,\dots,N,\ n \neq m$$

where $U_m'$ denotes the first derivative with respect to $X_m(\theta)$ and the arguments of $U_m$

[9] The utility function may be state dependent and may be an indirect utility function derived from optimal consumption decisions after $\theta$ has occurred in period 2. The model could also be formulated as a 2-period consumption model as in Ekern (1974).

[10] The existence of an equilibrium is assumed here. Oliver Hart and Dimitri Bertsekas have presented necessary and sufficient conditions for the existence of an equilibrium in a capital market model.

and $U'_m$ are understood.[11] An investor equilibrium for a given set of decisions by firms is characterized by the conditions in (9), (10), and (11), and by the market clearing conditions in (7).

### E. *Arbitrage between Spot and Forward Markets*

The interest rate parity theorem may be verified by rewriting $\rho_{mn}(\theta)\rho_{nj}(\theta)$ in (10) as $\rho_{mj}(\theta)$, solving (11) for $E_m(U'_m \cdot \rho_{mj}(\theta))$, and by substituting into (10) to obtain

$$E_m\{U'_m \cdot (r_{mj}(i_m/i_j) - \rho_{mn}(\theta)\rho_{nj})A_{nj}\} = 0$$

Then solving (11) for $E_m(U'_m \cdot \rho_{mn}(\theta))$ and substituting above, yields

$$E_m\{U'_m \cdot [r_{mj}(i_m/i_j) - \rho_{nj}r_{mn}(i_m/i_n)]A_{nj}\} = 0$$

The term in brackets does not involve $\theta$ and marginal utility is positive, so in equilibrium the term in brackets must equal zero. That term may be rewritten as

$$\rho_{nj} = r_{nj}i_n/i_j \tag{12}$$

since $r_{mj}/r_{mn} = r_{nj}$. The interest rate-parity theorem is thus verified and all gains due to arbitrage between spot and forward markets will be eliminated at an investor equilibrium. The arbitrage mechanism has not been introduced explicitly in this model but instead the two components of that form of arbitrage, borrowing or saving in a foreign currency and a matching forward contract, have been utilized. The significance of the interest rate-parity theorem will be considered in Section IV.

## II. Firm Production and Covering

### A. *Processes for Firm Decision Making*

The process by which firms make production and covering decisions has yet to be specified, and that process is subject to debate with firms being hypothesized to maximize their market value, to maximize the expected utility of their shareholders, or to maximize the preferences of their managers, for example. Leland (1974), Stiglitz, and Ekern and Wilson have demonstrated that in an incomplete capital market, value maximization does not in general result in Pareto optimal resource allocations so attention will be focused on the latter two processes.[12] The maximization of the expected utility of shareholders is certainly the most reasonable criterion for a firm, but firms do not in general know the preferences, probability assessments, and portfolio holdings of their shareholders. Although a firm does not know the characteristics of its shareholders, it can consider a hypothetical shareholder. Such a shareholder will necessarily have portfolio holdings that satisfy the conditions in (9), (10), and (11) for some utility function, some probability assessments, and some initial wealth. The firm may then determine how its decisions affect the optimal expected utility of a shareholder at his optimal portfolio holdings.

In order to investigate an investor's preference for the forward covering by the firm, consider a change $\Delta A_{nj}$ in covering at an investor equilibrium where the optimal decisions are denoted by a ($\hat{}$). The resulting change in the investor's return in (6) is

$$\rho_{mn}(\theta)(\rho_{nj}(\theta) - \rho_{nj})\Delta A_{nj}(\hat{t}_{mnj} - \hat{z}_{mn}) \tag{13}$$

Given such a change in covering by the firm, suppose that each investor alters his foreign currency holdings as follows:

$$\begin{aligned} \Delta y_{mj} &= (\hat{z}_{mn} - \hat{t}_{mnj})\Delta A_{nj}/i_j \\ \Delta y_{mn} &= -(\hat{z}_{mn} - \hat{t}_{mnj})\Delta A_{nj}\rho_{nj}/i_n \end{aligned} \tag{14}$$

[11] The condition in (10) for $t_{mnj}$ is equivalent to that for $t_{mjn}$, since $\rho_{nj} = 1/\rho_{jn}$ implies $\rho_{mn}(\theta)(\rho_{nj}(\theta) - \rho_{nj}) = \rho_{mj}(\theta)(\rho_{jn} - \rho_{jn}(\theta))/\rho_{jn}$. Consequently, an optimal solution can be taken to have $\hat{t}_{mnj} - \hat{t}_{mjn}$ as the investor's share of the net forward sales of currency $j$ against $n$.

[12] The criterion of value maximization also does not lead to optimal output levels that are independent of investor characteristics as do the other two criteria.

$$\begin{aligned}(15)\quad \frac{\partial E_m U_m}{\partial q_{nj}} = E_m \Bigg\{ U'_m \cdot \Bigg[ &\sum_k (\bar{z}_{mk} - \hat{z}_{mk}) i_m r_{mk} \frac{\partial V_k}{\partial q_{nj}} \\ + &\sum_k (\pi_k(\theta)\rho_{mk}(\theta) - i_m r_{mk} V_k) \frac{\partial \hat{z}_{mk}}{\partial q_{nj}} \\ + &\sum_{\substack{l \\ l \neq k}} \sum_k \rho_{ml}(\theta)(\rho_{lk}(\theta) - \rho_{lk}) A_{lk} \frac{\partial \hat{l}_{mlk}}{\partial q_{nj}} \\ + &\sum_k (i_k \rho_{mk}(\theta) - i_m r_{mk}) \frac{\partial \hat{y}_{mk}}{\partial q_{nj}} \\ + &\rho_{mn}(\theta)(\rho_{nj}(\theta) MR_{nj} - MC_{nj}) \hat{z}_{mn} \Bigg] \Bigg\}\end{aligned}$$

Adding the returns from (13) and (14) indicates that the return $X_m(\theta)$ in (6) in each state $\theta$ is then unchanged by the change in the firm's covering, since the change in the return on forward contracts plus profits is exactly offset by the change in the return on foreign currency borrowing and lending. If the investor's alterations in his currency purchases are feasible, the distribution of return optimal before the change can be duplicated after the change in covering. Then no investor has an incentive to alter his shareholding or his share of the forward contracts, so the value of the firm (and all other firms) must remain the same. To show that the investor's alterations in his currency purchases are feasible, determine the net change in the home currency cost of the purchases in (14) which is

$$r_{mj}\Delta y_{mj} + r_{mn}\Delta y_{mn} = (\hat{z}_{mn} - \hat{l}_{mnj})\Delta A_{nj}(r_{mj}/i_j - \rho_{nj} r_{mn}/i_n)$$

The arbitrage relationship $\rho_{nj} = r_{nj} i_n / i_j$ in (12) then implies that this term is zero. This holds for all investors at any output levels for firms, so the investor is able to exactly offset or duplicate the covering of any firm. Consequently, the value of a firm is independent of its covering, and all investors are indifferent to the covering of any firm. This homemade covering result is stated as Proposition 1.

PROPOSITION 1: *At an investor equilibrium for given output levels of firms, all investors are indifferent to the covering decisions of any firm, and hence the value of a firm is independent of the covering by any firm.*[13]

In an incomplete capital market, an investor is not able to duplicate the return possibilities generated by the productive activities of firms, and hence, investors are not globally indifferent to the production levels of firms. The preferences of an investor for the exports $q_{nj}$ may be determined by differentiating his optimal expected utility with respect to $q_{nj}$ to obtain[14] equation (15). The marginal revenue in currency $j$ is:

$$MR_{nj} = p'_{nj}(q_{nj})q_{nj} + p_{nj}(q_{nj})$$

[13] The same result obtains for firm borrowing and lending in foreign currencies, which justifies leaving such possibilities out of the model. Since the firm has no guide as to its covering decisions, the values of $A_{nj}$ will be taken as nonzero in the following analysis.

[14] The output of firm $n$ is assumed not to affect the profit of any other firm, so that there are no industry effects. Such industry effects are considered by Ekern (1975).

The marginal cost in currency $n$ is

$$MC_{nj} = \partial C_n(q_n)/\partial q_{nj}$$

and the value of the firm is assumed to be differentiable with respect to $q_{nj}$. The second, third, and fourth terms of (15) are zero at an investor equilibrium from (9), (10), and (11).

The analysis of the firm's decisions will be conducted at an *ex post* equilibrium in which investors already hold optimal portfolios given the decisions of firms.[15] Investors may be thought of as having gone through a series of portfolio revisions until arriving at the portfolio holdings ($\bar{z}_{mn}$, $\bar{y}_{mn}$, $\bar{l}_{mnj}$) that are optimal, and then the investor does not wish any further revision. Thus $\hat{z}_{mn}=\bar{z}_{mn}$, $\hat{y}_{mn}=\bar{y}_{mn}$, and $\hat{l}_{mnj}=\bar{l}_{mnj}$. The first term in (15) is then equal to zero, and the investor's preference for a (local) change in $q_{nj}$ is given by

$$(16) \quad \frac{\partial E_m U_m}{\partial q_{nj}} = E_m\{U'_m \cdot \rho_{mn}(\theta)(\rho_{nj}(\theta)MR_{nj} - MC_{nj})\}\hat{z}_{mn}$$

As expressed in (16) the preferences of an investor for changes in output depend on the investor's utility function, probability assessments, and portfolio holdings. The investor equilibrium, however, reflects these investor characteristics and establishes prices for the risks that are traded in the foreign exchange markets. For example, scaling expected marginal utility $E_m U'_m$ in (11) to equal one and using (12) yields

$$(17) \quad E_m\{U'_m \cdot \rho_{mn}(\theta)\} = \rho_{mn}$$

This holds for all investors, so $\rho_{mn}$ may be interpreted as the market price for a claim to one unit of currency $n$ in period 2. The role of the forward market in taking into account divergent investor characteristics may be seen by writing (17) as

$$E_m(\rho_{mn}(\theta)) - \rho_{mn} = -\,cov_m(U'_m, \rho_{mn}(\theta))$$

where $cov_m$ is the covariance of $U'_m$ and $\rho_{mn}(\theta)$. The left side is the difference between the expected future spot rate and the forward rate. If the investor believes that the expected future spot rate will be greater (less) than the forward rate, the investor will adjust his portfolio so that the covariance between marginal utility and the future spot rate will be negative (positive). For example, if the left side is positive, the investor will choose a portfolio that has positive net claims to currency $j$ through share ownership, savings, and forward sales. A similar interpretation may be given to the condition in (10).

This market price may be used to further analyze investor preferences in (16). Substituting $E_m\{U'_m \cdot \rho_{mn}(\theta)\}$ from (11) and similarly $E_m\{U'_m \cdot \rho_{mj}(\theta)\}$ from (11) into (16) yields

$$(18) \quad \frac{\partial E_m U_m}{\partial q_{nj}} = E_m\{U'_m\}\rho_{mn}(\rho_{nj}MR_{nj} - MC_{nj})\hat{z}_{mn}$$

Since expected marginal utility is positive as is $\rho_{mn}$, the sign of the expression in (18) depends on the marginal "fully covered profit" ($\rho_{nj}MR_{nj}-MC_{nj}$) on the firm's exports to country $j$ and the sign of $\hat{z}_{mn}$. This is true for all investors, so all shareholders ($\hat{z}_{mn}>0$) will unanimously prefer a (local) increase (decrease) in the exports to country $j$ if and only if the marginal fully covered profit is positive (negative). Shareholders will prefer no change in the outputs of a firm when

$$(19) \quad \rho_{nj}MR_{nj} - MC_{nj} = 0, \quad j = 1, \ldots, N$$

The firm may thus serve the interests of its shareholders by acting as a multiproduct firm using the forward exchange rates as planning equivalents. This establishes the

[15] Leland (1974) and Ekern (1974) have also used *ex post* equilibrium analysis. Radner has considered *ex ante* equilibrium analysis.

following shareholder unanimity and production planning proposition.

PROPOSITION 2: *All shareholders have unanimous preferences for (small) changes in the output levels of firms, and the firm may maximize the expected utility of its shareholders by acting as a multiproduct, profit-maximizing firm using the forward exchange rates as planning equivalents.*[16] *The export levels that equate fully covered marginal revenue to marginal cost are Pareto optimal.*[17]

The production levels are thus independent of the characteristics of investors except to the extent that those characteristics establish the forward exchange rates. This result obtains because investors are able to offset any exchange rate risk created by the exports of the firms by trading in forward exchange markets and/or by borrowing and lending in foreign currencies. For example, if the firm has export revenue $R_{nj}$ in currency $j$, the investor can offset that exchange rate risk by borrowing a quantity $R_{nj}\hat{z}_{mn}/i_j$ of currency $j$ to cover the exchange rate risk generated by exports. In period 2 the investor's share of the firm's exports will effectively cover the loan obligation. Proposition 1 reflects this homemade covering effect which implies that the firm can ignore the exchange rate uncertainty in planning its operations and may use the forward rate as a "planning equivalent."

Another implication of the homemade covering result of Proposition 1 is that production may be planned independently of any capital market considerations. The results of Leland (1974) and Ekern (1974), for example, indicate that production will in general be a function of the value of the firm. In the model considered here, the production levels are independent of the value of the firm (except in that the forward exchange rates and firm values are determined simultaneously in equilibrium) because the existence of the forward exchange markets permits investors to efficiently price the exchange risks.

The process by which the firm determines its optimal output levels may be thought of as a voting process in which the firm submits its planned output levels to its shareholders for ratification. Given any output plan, shareholders will be unanimous with respect to their preferences for small changes in the plan, and the process will terminate when (19) is satisfied. Such a voting process is not necessary however because a firm can consider any hypothetical shareholder and can predict the response of that shareholder and hence the responses of all shareholders. To do so, the firm need not know anything about shareholders except that they maximize the expected utility of their portfolio return. The

[16] Proposition 2 is an *ex post* unanimity result (see Leland 1973, 1974, Ekern and Wilson, Ekern 1974, and Radner) that obtains because for each state $\theta$ the marginal returns with respect to a firm's decisions may be expressed as a linear combination of the returns on foreign currency holdings. The unanimity result is *ex post* in the sense that it holds at the optimal portfolio for investors. Proposition 1 is stronger than Proposition 2 because the existence of a forward market or the opportunity for foreign currency borrowing and lending allows the investor to create the same opportunities that a firm can create through its covering. Investors are not assumed to be able to create the same pattern of return as can be generated by the output decisions of firms. If the stronger assumption is made that investors can duplicate those patterns of return, unanimity results may be obtained for *ex ante* shareholders. The same production planning result obtains under these assumptions.

[17] To demonstrate that the output levels are Pareto optimal, consider the problem of maximizing the expected utility of one investor subject to constrained levels of expected utility for other investors. Thus, for investor $m$ the program is max $E_mU_m(X_m(\theta))$ subject to $E_lU_l(X_l(\theta))=U_l^0$, $l=1,\ldots,N$, $l\neq m$, where $U_l^0$ is a fixed level of expected utility. The necessary condition for Pareto optimality for $q_{nj}$, for example, is

$$E_m\left\{U_m'\cdot\frac{\partial X_m(\theta)}{\partial q_{nj}}\right\}-\sum_{l\neq m}\lambda_{ml}E_l\left\{U_l'\cdot\frac{\partial X_l(\theta)}{\partial q_{nj}}\right\}=0$$

where $\lambda_{ml}$ is a multiplier associated with the constrained expected utility of investor $l$. From (9), (10), (11), (18), and the analysis of (13) and (14), this condition is satisfied for every investor, at an *ex post* equilibrium.

determination of the optimal output levels thus is an efficient process that respects the privacy of investors.

An alternative method of determining output levels would be to explicitly recognize the separation between the owners and the managers of firms. The managers of firms may have profit objectives reflecting their own preferences and probability assessments, and even if they do wish to act in the best interests of shareholders, the management may seek to do so by accepting a criterion such as maximizing the expected utility of profit. The criterion of maximizing the expected utility of profit for some utility function and some probability assessments has been frequently used in the literature[18] as a surrogate for the maximization of the expected utility of investors' return, and if firms are not assumed to have performed the previous analysis, such a criterion might seem reasonable. Furthermore, the management of a firm might believe that it has more accurate knowledge of the likelihood of various exchange rate levels obtaining, for example, and thus might use its own probability assessments in making its decisions.

Suppose that firm $n$ has chosen a strictly concave, measurable utility function $\mathcal{U}_n$ and a probability distribution over the states of nature with expectation denoted by $\mathcal{E}_n$ and seeks to maximize

$$\mathcal{E}_n \mathcal{U}_n(\pi_n(\theta))$$

with respect to $(q_{nj}, A_{nj})$, $j = 1, \ldots, N$. The necessary optimality conditions are

$$(20) \quad \frac{\partial \mathcal{E}_n \mathcal{U}_n}{\partial q_{nj}} = \mathcal{E}_n\{\mathcal{U}_n' \cdot (\rho_{nj}(\theta) MR_{nj} - MC_{nj})\} = 0, \quad j = 1, \ldots, N$$

$$(21) \quad \frac{\partial \mathcal{E}_n \mathcal{U}_n}{\partial A_{nj}} = \mathcal{E}_n\{\mathcal{U}_n' \cdot (\rho_{nj} - \rho_{nj}(\theta))\} = 0, \quad j = 1, \ldots, N, j \neq n$$

[18] See the author and Leland (1972), for example.

Solving (21) for $\mathcal{E}_n\{\mathcal{U}_n' \cdot \rho_{nj}(\theta)\}$ and substituting into (20) yields

$$\mathcal{E}_n\{\mathcal{U}_n'\} \cdot (\rho_{nj} \cdot MR_{nj} - MC_{nj}) = 0, \quad j = 1, \ldots, N$$

Since expected marginal utility is positive, the term in parentheses must be zero and is thus identical to (19).[19] Consequently, a firm that maximizes the expected utility of profit for any utility function and probability assessments will produce such that all shareholders prefer those production levels. This is again a form of a homemade covering result, since the firm may compensate for the exchange rate risk created by exportation by selling (or buying) foreign currencies forward. The amount of such forward transactions will depend on the expectations and the utility function used, but since investors are indifferent to the forward exchange transactions of firms, the dependence of the covering decisions on the expectations and preferences of management does not affect investors' expected utility.

Consequently, the firm may act as a maximizer of expected utility of profit or equivalently may act to directly maximize the expected utility of a shareholder, and the resulting output levels will be Pareto optimal and will maximize the expected utility of all shareholders. Both of these processes are efficient in the sense that they require no communication between firms and investors, and both respect the privacy of investors. The equivalence of the two processes in terms of the output levels of firms obtains because of the opportunity to trade foreign currency risks in a market. While the level of such covering by firms may differ for the two processes, the value of a firm will be identical, since that value depends on the real pro-

[19] This result has also been obtained by Wilfred Ethier. The existence of positive output levels satisfying (20) and (21) is assumed.

ductive decision of the firm and not on the covering of the exchange rate risks as indicated in Proposition 1.

### B. *Firm Valuation*

While the firm will use the forward exchange rate as a planning equivalent, the firm's profit will be uncertain unless the firm fully covers its exchange risk. The value of the firm may be determined by eliminating the future spot rates from (9) by using (10) and (11) to obtain

$$(22)\quad E_m\{U'_m\}\left[\rho_{mn}\left(\sum_j \rho_{nj}R_{nj} - C_n(\hat{q}_n)\right) - i_m r_{mn} V_n\right] = 0$$

where $\hat{q}_n$ is the vector of outputs satisfying (19). The term in brackets does not involve $\theta$, so the value of the firm is

$$(23)\quad V_n = \left(\sum_j \rho_{nj}R_{nj} - C_n(\hat{q}_n)\right) \Big/ i_n$$

which is the fully covered profit of the firm discounted at the risk-free rate of interest. The value is independent of the level of covering as previously indicated, but the value of the firm depends on the forward exchange rates as do the output levels. The production planning condition in (19) may be used to eliminate the forward exchange rates from (23) yielding

$$(24)\quad V_n = \sum_j (R_{nj}(MC_{nj}/MR_{nj}) - C_n(\hat{q}_n))/i_n$$

The value of the firm is thus determined completely by its revenue and cost functions. The term $(MC_{nj}/MR_{nj})$ converts the currency $j$ revenue $R_{nj}$ to currency $n$, and the equivalence between this conversion and the usual conditions for a multiproduct firm is evident by noting that for a firm that does not export $(j=n)$, $MC_{nn}=MR_{nn}$. The profit is thus just the difference between revenue and cost. If firm $n$, for example, sells in only country $j$ and the market in that country is competitive $(R_{nj}/MR_{nj}=\hat{q}_{nj})$, the value of the firm is

$$V_n = \hat{q}_{nj}(MC_{nj} - AC_n)/i_n$$

where $AC_n=C_n(\hat{q}_{nj})/\hat{q}_{nj}$ is average cost. The value of the firm will be positive if the marginal cost exceeds average cost at the optimal output, which is the usual valuation rule for a competitive firm producing an output such that price equals marginal cost.

### C. *Comparative Statics of Trade Policies*

Within the context of the model a variety of governmental policies can be analyzed. For example, a tariff $T_j$ imposed by country $j$ on imports reduces the revenue of an exporter to $(p_{nj}(q_{nj})-T_j)q_{nj}$ and the marginal revenue to $(MR_{nj}-T_j)$. Such a tariff will reduce the imports of that country in the same manner as in a deterministic model. Such a tariff can, of course, be offset by the exporting country with a per unit subsidy of $\tau_j\rho_{nj}=T_j$ from country $n$. A lump sum subsidy or import license fee does not affect the quantity of imports or exports directly but does reduce the value of the firm by the amount of the fee, and if the cost of the import license is great enough, a firm will not export. Similarly, the exports of a firm are independent of the profits tax rate in the firm's home country. Further analysis will be left to the reader.

## III. The Role of Forward Markets

With the results of Proposition 2, one can compare fixed and flexible exchange rate systems. A fixed exchange rate system may be considered as a special case of a flexible exchange rate system in which all market participants have degenerate probability assessments with $\rho_{mn}(\theta)=\rho_{mn}=r_{mn}(i_m/i_n)$ for all $\theta$, $m$, and $n$. The outputs resulting with a fixed exchange rate system will be those obtaining with a flexible ex-

change rate system as specified in (19), so the two systems may be considered to be equivalent in the context of this model. This equivalence is insured by the existence of forward markets or by the opportunity to borrow and lend in foreign currencies. Either permits the firm to adopt one of the processes in Section II and hence to use the forward rate for planning purposes.

If forward exchange markets do not exist and investors are prohibited from borrowing and lending in foreign currencies, the level of trade will be the same as that given in Section II if there exists a firm in each country with a nonzero value that sells only in that country. That firm faces no uncertainty, and the necessary portfolio optimality conditions for such firms for an investor in country $m$ are from (9)

$$(25)\quad E_m\{U'_m\cdot(\rho_{mn}(\theta)(R^k_{nn}-C^k_n) - r_{mn}i_mV^k_n)\} = 0, \qquad n = 1,\ldots,N$$

where the superscript $k$ denotes the riskless firm. Solving this for $E_m\{U'_m\cdot\rho_{mn}(\theta)\}$ and $E_m\{U'_m\cdot\rho_{mj}(\theta)\}$ and substituting into (16) yields

$$(26)\quad \frac{\partial E_mU_m}{\partial q_{nj}} = [E_mU'_m]\left(MR_{nj}\frac{r_{mj}i_mV^k_j}{R^k_{jj}-C^k_j} - MC_{nj}\frac{r_{mn}i_mV^k_n}{R^k_{nn}-C^k_n}\right)\hat{z}_{mn}$$

For an investor in country $n$ the condition in (25) implies that the value of the riskless firm is the present value of profits or $V^k_n=(R^k_{nn}-C^k_n)/i_n$. Using this condition for $n$ and $j$ in (26) yields

$$(27)\quad \frac{\partial E_mU_m}{\partial q_{nj}} = [E_mU'_m](MR_{nj}r_{mj}i_m/i_j - MC_{nj})(r_{mn}i_m/i_n)\hat{z}_{mn}$$

Since $(r_{mj}i_m/i_j)$ would equal the forward exchange rate if it existed, the preferences of all shareholders for changes in $q_{nj}$ are the same as those given in (18). Consequently, the opportunity to trade in the shares of a riskless firm in each country provide sufficient patterns of returns for unanimity to result and for firms to use the equivalent forward rate for the planning of exports.

The opportunity to trade in the securities of a riskless firm in each country provides the same patterns of returns in the economy as may be created by borrowing and lending in the currency of that country. Consequently, if there is no riskless firm in a country, but borrowing and lending in that country are permitted, the same results as those with a forward market obtain. The forward exchange market, the borrowing and lending markets in foreign currencies, and securities trading of the shares of riskless firms are thus redundant in the sense that if only one existed, the resulting equilibrium would be the same. The interest rate-parity theorem is essentially a statement of this redundancy.

Even if none of these three market opportunities is available to investors, unanimity may result regarding the levels of a firm's exports. As an example, suppose that there is a firm in country $m$ that exports only to one country, $j$. The conditions in (9) and (18) for an investor in country $m$ who may not make forward transactions, foreign borrowing or lending, or purchase shares of riskless foreign firms are

$$(9a)\quad E_m\{U'_m\cdot[\rho_{mj}(\theta)R_{mj} - C_m - i_mV_m]\} = 0$$

$$(18a)\quad \frac{\partial E_mU_m}{\partial q_{mj}} = E_m\{U'_m\}[(MR_{mj}/R_{mj})\cdot(C_m+i_mV_m) - MC_{mj}]\hat{z}_{mm}$$

The term in brackets in (18a) is independent of $\theta$, and all shareholders in country $m$ prefer an increase (decrease) in $q_{mj}$ if that term is positive (negative). All such

investors are unanimous with respect to their output preferences, but the preferred output level will be different from that in (19). The output level in this case will depend on the value of the firm.[20,21]

In reality, there are investors who invest in securities of firms but who do not consider trading in foreign currencies. The indifference and unanimity results still obtain for the shareholders in the former group but not those in the latter group. For the latter group the sign of the term in (16), for example, depends on the preferences and probability assessments of that investor, and hence production planning is complicated by the same issues that arise with public goods.

With this type of market segmentation shareholder unanimity may still result for certain firms. Consider the above example of the firm in country $m$ that exports only to country $j$. If there are some investors in country $m$ who do trade in a forward exchange market or its equivalent, shareholders in this second group will prefer an increase in output if and only if $\rho_{mj}MR_{mj} - MC_{mj} > 0$ from (18). From (23) the value of the firm will be $V_m = (\rho_{mj}R_{mj} - C_m)/i_m$ which implies in (18a) that shareholders in the first group will be in agreement with the shareholders in the second group. If the output satisfying (19) is produced, then (18a) is zero, and all shareholders will achieve their optimal expected utility.

Another type of market segmentation will exist if investors in one country do not consider trading in the security markets in other countries, so that the investor equilibrium conditions in (9) hold only for $n = m$. If those investors trade in a forward market or its equivalent, all investors are again indifferent to the covering decisions of firms, and shareholders in a firm are unanimous with respect to their preferences for changes in output. The value of the firms and the forward exchange rates are also unchanged. The only essential effect of such segmentation is that these results must be developed by examining the capital market in each country instead of a single market. If the investors also do not trade in the forward exchange market or its equivalent, then the unanimity and indifference results may not obtain, in general.[22]

## IV. Conclusions

Forward exchange markets or their equivalents permit investors to perform their own covering of the uncertain returns from share ownership, and consequently, firms are able to plan their outputs using the forward exchange rate as a planning equivalent. The output levels determined by equating fully covered marginal revenue and marginal cost will be unanimously supported by all shareholders and will maximize their expected utility. These output levels may be achieved without the firm knowing any characteristics of its shareholders, and thus, the planning process respects the privacy of shareholders. The resulting value of the firm will be the present value of its fully covered profit, which is independent of the covering of the firm because all investors are indifferent to that forward covering. These results must be qualified to the extent to which there are imperfections in the capital or foreign exchange markets or if there are additional forms of uncertainty that cannot be efficiently allocated in competitive markets.

[20] This result is analogous to that obtained by Leland (1974) and Ekern (1974).

[21] This example may be generalized for firms that may export to any number of foreign countries. If the system of equations in (9) may be solved for $E_m[U'_m\rho_{mn}(\theta)]$, $n=1, \ldots, N$, for each investor, unanimity will result, but in general the output of a firm will depend on the revenues, costs, and values of all firms. With one firm in each country a necessary and sufficient condition for this is that the inverse exist for the $N \times N$ matrix with elements $R_{ij}$, $j=1, \ldots, N$, $i=1, \ldots, N$, $j \neq i$, and $R_{ij} - C_i$ for $i=j$.

[22] Unanimity may still result if, for example, the condition in fn. 21 is satisfied.

In the absence of such imperfections the forward exchange markets duplicate the borrowing and lending markets as indicated by the interest rate-parity theorem. The presence of a riskless firm in each country also provides equivalent market opportunities to guarantee unanimity. To achieve unanimity, it is not necessary that investors trade in foreign capital markets if they trade in the forward exchange market or its equivalent. Even if investors do not trade in forward markets or their equivalents, unanimity may result although the output levels may be different as demonstrated in the previous section.

## REFERENCES


**D. P. Baron**, "Price Uncertainty, Utility, and Industry Equilibrium in Pure Competition," *Int. Econ. Rev.*, Oct. 1970, *11*, 463–80.

**D. P. Bertsekas**, "Necessary and Sufficient Conditions for Existence of an Optimal Portfolio," *J. Econ. Theory*, June 1974, *8*, 235–47.

**P. A. Diamond**, "The Role of a Stock Market in a General Equilibrium Model with Technological Uncertainty," *Amer. Econ. Rev.*, Sept. 1967, *57*, 759–76.

**S. Ekern**, "Some Aspects of Firms' Decision Making in an Economy with Incomplete Capital Markets," *Swedish J. Econ.*, Mar. 1974, *76*, 117–30.

———, "On the Theory of the Firm in an Economy with Incomplete Markets: An Addendum," *Bell J. Econ.*, Spring 1975, *6*, 388–93.

——— **and R. Wilson**, "On the Theory of the Firm in an Economy with Incomplete Markets," *Bell J. Econ.*, Spring 1974, *5*, 171–80.

**W. Ethier**, "International Trade and the Forward Exchange Market," *Amer. Econ. Rev.*, June 1973, *63*, 494–503.

**S. Grassman**, *Exchange Reserves and the Financial Structure of Foreign Trade*, Westmead 1973.

**H. G. Grubel**, *Forward Exchange, Speculation, and the International Flow of Capital*, Stanford 1966.

**O. D. Hart**, "On the Existence of Equilibrium in a Securities Model," *J. Econ. Theory*, Nov. 1974, *9*, 293–311.

**H. E. Leland**, "Theory of the Firm Facing Uncertain Demand," *Amer. Econ. Rev.*, June 1972, *62*, 278–91.

———, "Production Theory and the Stock Market," *Bell J. Econ.*, Spring 1974, *5*, 125–44.

———, "Capital Asset Markets, Production, and Optimality: A Synthesis," tech. rep. no. 115, Inst. Mathematical Stud. Soc. Sci., Stanford Univ., Dec. 1973.

**R. Radner**, "A Note on Unanimity of Stockholders' Preferences Among Alternative Production Plans," *Bell J. Econ.*, Spring 1974, *5*, 181–84.

**J. E. Stiglitz**, "On the Optimality of the Stock Market Allocation of Investment," *Quart. J. Econ.*, Feb. 1972, *86*, 25–60.

# Related Market Conditions and Interindustrial Mergers

*By* M. L. Greenhut and H. Ohta*

William Baumol and David Bradford once introduced a paper in this *Review* by suggesting that its subtitle could be: *The Purloined Proposition or The Mystery of the Mislaid Maxim.* They proposed that the results contained in their paper had already appeared often in the literature, but that these results would still surprise many readers who tend to consider them at variance with ideas they had long accepted. The same caveat applies in many ways to the present paper. However, the literature which contends (as we do below) that important benefits stem from vertical integration has been predicated essentially on analysis of bilateral monopoly. It was, perhaps, this condition that prompted Fritz Machlup and Martha Taber to conclude their review of the literature by, in effect, rejecting the *Maxim* in question. Possibly this *Maxim*, even as repeated by S. Y. Wu, has been and continues to be *Mislaid* because proofs have been sketchy, chiefly diagrammatic, and the subject joined with that of bilateral monopoly. The upshot has been that a vital part of the *Maxim* has been lost, namely that the benefits stemming from vertical merger of successive monopolies dovetail with the effects of having perfect competition on the higher stage of production rather than monopoly. This particular identification, perhaps even more so than the generality of our proofs, nurtures our hope that the *Maxim* in question will not be permanently mislaid. We consider this *Maxim*, as set forth below, to be especially vital in this period of energy shortage and attacks on big business.

Our paper considers two types of related firms: producers of goods who have no monopsony power over input suppliers, and a monopolist input supplier.[1] The two theorems which will be deduced here are (I) *the price charged by the monopolistic supplier of an input is not affected by the market structure in the market for the final good, be it perfectly competitive, monopolistic, etc.*, and (II) *merger or collusion between the input supplier and the final good producer brings about lower prices, greater output and sales, and greater profits to the merged or colluding firms—a welfare gain.*[2]

In Section I we present our theorems for

* Professor, Texas A&M University, and associate professor, Aoyama Gakuin University, respectively. We wish to thank Ming-jeng Hwang for critiques and suggestions and Martin Bronfenbrenner who pointed out the intimate relation of the present work to that of Augustin Cournot. This paper is based in part on research funded by the National Science Foundation.

[1] Our analysis relates to all connected stages of production, e.g., raw materials to intermediate products to final products, consignor to consignee, inputs of labor in producing goods, etc. Simplicity in reference is gained, however, by conceiving generally in this paper of two particular connected stages of production: A) production of final goods, and B) production of the raw materials used as inputs in producing the final goods.

[2] In a sense, this second theorem is related to Cournot's theorem on production with joint inputs. However, some significant differences in analytical framework exist between the two. To begin with Cournot assumes that $n$ kinds of raw materials, e.g. copper and zinc, are *jointly* used to produce a commodity, e.g. brass. His problem was to analyze the impact on price and quantity produced of a *horizontal* merger or collusion between the producers of the raw materials when the final commodity is sold under conditions of perfect competition. In contrast, our model will include among others the case where all producers are monopolistic, and our problem is to investigate the impact of *vertical* collusion between firms in different stages of production and/or distribution.

the two polar cases—perfect competition and monopoly—in the final good market structure. Section II extends our theorems to a Cournot-oligopoly final good market structure. Section III establishes the basic welfare results for varying as well as constant marginal cost situations and demonstrates the generality of our results by illustrating their applicability to such inputs as transportation services. Section IV concludes the paper with a discussion of some antitrust implications of the theory.

## I

Consider a final good market demand function $f$ of the form:

$$(1)\quad p=f(q),\quad f'=\frac{df(q)}{dq}<0,\quad f(0)<\infty$$

where $q$ stands for the quantity demanded at whatever price $p$ is charged in that market. The correspondent to $f(q)$ may then be defined as:[3]

$$(2)\quad g(q)=f(q)+f'q,\qquad f'=\frac{df(q)}{dq}$$

$$=f(q)\left(1-\frac{1}{e_f}\right),\quad e_f=-\frac{dq}{df(q)}\frac{f(q)}{q}$$

The graphs of $f(q)$, $g(q)$, and $h(q)$, i.e., the correspondent to $g(q)$, are drawn respectively in Figure 1a, which assumes *for simplicity* a linear demand. Our analysis, however, is not restricted to this linearity assumption, as will shortly be demonstrated.

Three situations warrant analysis initially, namely (i) where perfect competition prevails in the final good (commodity) market and the input supplier is a simple monopolist; (ii) simple monopoly prevails in each market; and (iii) the monopolists in the two related industries collude. Later, in Section II, we shall examine as well the market where the final good is sold under imperfectly competitive conditions.

CASE (i): Consider first the case where the producer of the final good, firm $A$, is a perfect competitor. His profit maximizing calculus requires

$$(3)\qquad p=MC=c+r$$

where $c$ relates to the original factor inputs used by firm $A$ (say, labor) and $r$ to the quantity of raw materials purchased from the input supplier, firm $B$, per unit of output $q$. Without loss of generality, $c$ and $r$ are assumed to be constant for the time being, an assumption relaxed later in Section II of the paper. Firm $A$'s cost $r$ constitutes firm $B$'s revenue. Combining (1) with (3) provides the revenue function of $B$ per unit output $q$ on sales to $A$, i.e.:

$$(4)\qquad r=f(q)-c$$

Defined as such, (4) represents the average revenue of firm $B$ in terms of $q$, *not* the average revenue (or price) it receives *per unit of raw materials* sold to firm $A$ (and to other atomistic firms in the final good industry).[4] The marginal revenue for firm $B$, $MR_B$, is then:

$$(5)\qquad MR_B=g(q)-c$$

$$=f(q)\left(1-\frac{1}{e_f}\right)-c$$

[3] The correspondent $g(q)$ is, in effect, the marginal revenue function of the final good producer if and only if monopoly prevails in that market. However, if perfect competition prevails in that market, the correspondent $g(q)$ constitutes the marginal revenue function for the monopolistic input supplier, indirectly and in part, but of course not for the competitive producer of the final good (see equation (3)). Incidentally, the concept of *correspondent* used herein should not be confused with Joan Robinson's correspondent which relates to a tangent to an average revenue function.

[4] Let the input material supplied by firm $B$ be represented by $M$ and the price of $M$ by $p_M$. Then the average revenue of a unit of $M$ can be specified as $p_M=(f(q)-c)q/M$. The average revenue per unit of $q$, specified in the text, is then readily derived as $r=(p_MM)/q=f(q)-c$. Note further that a constant $r$ requires constant proportions between $M$ and $q$ for any $p_M$ parametrically given to firm $A$; as noted above in the text, this fixed relation will be relaxed later in the paper.

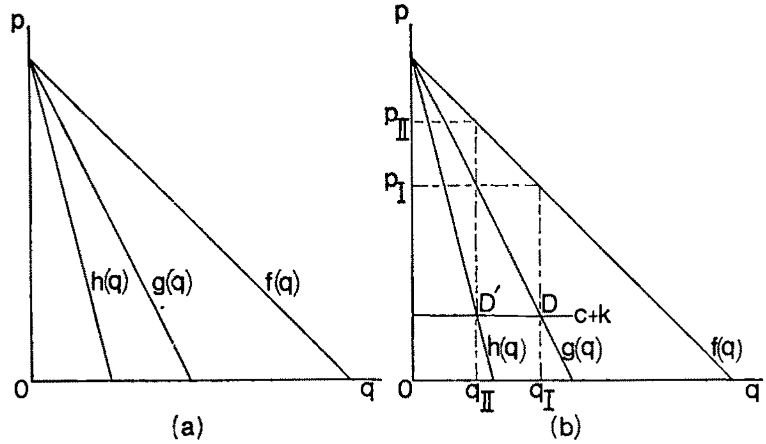



FIGURE 1

where the marginal revenue relates to an extra unit of output $q$, not input.

Profit maximization by firm $B$ requires equating $MR_B$ with its marginal cost of production $k$, also assumed to be constant for the moment. Thus:

$$(6) \qquad f(q)\left(1 - \frac{1}{e_f}\right) - c = k$$

or

$$g(q) = c + k$$

Associated with this equilibrium is output $q = q_I$ depicted in Figure 1b, which is produced under conditions of perfect competition in the commodity market and monopoly in the raw material market.[5] Correspondingly, the price of the final good is $p = p_I$ in Figure 1b, the average revenue per unit of $q$ earned by firm $B$ is $r = p_I - c$, so the input supplier's profit on his production for firm $A$ is $\pi_B = [p_I - (c+k)]q_I$, while firm $A$ is covering costs with $\pi_A = 0$. Total profits in Case (i) are $\pi = \pi_B + \pi_A = [p_I - (c+k)]q_I$, an area easily observed in Figure 1b.

A mathematical summary of the preceding analysis helps shed light on certain important relationships. In particular, note that Case (i) consists of three fundamental equations, (1), (3), and (6), in three unknowns, $p$, $q$, and $r$. Equation (1), recall, represents the commodity market demand function, equation (3) the equilibrium condition for the competitive producers in that market, and equation (6) the equilibrium condition in the market of the monopolistic supplier of input materials. These equations may be solved simultaneously for $p$, $q$, and $r$ at the equilibrium point $D$, as depicted in Figure 1b.

CASE (ii): Consider now a *monopolist* final good producer who is subject to the demand function given as (1). His profit maximizing calculus, i.e. $MR_A = MC_A$, provides:

$$(7) \qquad g(q) = c + r$$

which is the counterpart equilibrium condition to (3). Equation (7), however, also establishes the average revenue function of the monopolistic input supplier firm $B$ in the form:

$$(8) \qquad r = g(q) - c$$

Firm $B$'s profit maximizing condition $MR_B = MC_B$ may, in turn, be obtained from (8) as:

$$(9) \quad g(q)\left(1 - \frac{1}{e_g}\right) - c = k, \; e_g = -\frac{dq}{dg}\frac{g(q)}{q}$$

[5] The amount $q = q_I$ is the aggregate supply by the $n$ competitive producers of the final product.

or $$h(q) = c + k$$

where $h(q)$ represents the correspondent to $g(q)$.[6] (And see the curve $h(q)$ in Figure 1.)

Equation (9) defines the equilibrium condition for firm $B$ stemming from a monopolized commodity market. This equilibrium is represented by point $D'$ in Figure 1b. Accordingly, the equilibrium output is $q=q_{II}$, the equilibrium price of the raw material package is $r=p_I-c$, firm $B$'s profit is $\pi_B=[p_I-(c+k)]q_{II}$, the price the final good producer (firm $A$) charges his buyers is $p=p_{II}$, and his profit is $\pi_A=(p_{II}-p_I)q_{II}$. The sum of the two independent monopolists' profits amounts to $\pi=\pi_B+\pi_A=[p_I-(c+k)]q_{II}$.

Figure 1b shows that output produced under the conditions of Case (i) is greater than that of Case (ii), i.e., $q_I>q_{II}$, and profits are also greater in Case (i) than Case (ii). Moreover, Figure 1b implies that the equilibrium input price is the same, i.e., $r=p_I-c$ in each case. *The input price charged by the monopolistic supplier is, therefore, invariant to the final good market structure*, be it either monopolistic or perfectly competitive. Moreover, it will be shown that the same conclusion holds also for imperfectly competitive final good markets. Section II will formally establish these relationships.

Though the equation system for Case (ii) has the same three unknowns, $p$, $q$, and $r$, it contrasts sharply from that of Case (i) by consisting of equations (1), (7), and (9) rather than (1), (3), and (6). As before, equation (1) sets forth the demand function, and equations (7) and (9) serve as counterparts to (3) and (6), respectively providing the average revenue and profit maximizing conditions for the input supplier, firm $B$.

CASE (iii): The greater profits available under Case (i) conditions compared to those of Case (ii) would induce the monopolistic supplier of inputs, firm $B$, to collude with the monopolistic producer of products, firm $A$. The demand function given by equation (1) would then become the relevant average revenue function for the integrated monopolist, and hence (2) provides the corresponding marginal revenue (alternatively $g(q)$ in Figure 1). This $MR$ would then be equated directly with the relevant marginal costs of producing the final product (which are related to the value-added of the final good and that of its input), i.e., $c+k$. Equilibrium price and output therefore apparently dovetail with that of Case (i); i.e., $q=q_I$ in Figure 1. The pooled profits $\pi$ for the integrated monopolist are, in turn, $\pi=\pi_B+\pi_A=[p_I-(c+k)]q_I$, with the profit resulting from collusion being the same as in Case (i).[7] Since profit in Case (i) exceeds that of Case (ii), it follows that profit in Case (iii) is also greater than that of Case (ii). Collusion pays from the standpoint of firms operating in related markets. Collusion also offers advantages to society since the final commodity price is lower and the output produced greater than otherwise. But the relations sketched above are based on the linear demand assumption; need exists for formal proof of these relations to include the more general case of non-linear demand.

A proof is available which requires two basic constraints on $f$ in addition to the requirements that $f'$, $g'$, $h'<0$ and $f(0)>0$: namely, 1) the demand function $f$ is limited to yield a correspondent $g$ such that the horizontal lengths of the two curves are in fixed proportion, as is the case when $f$ is of the algebraic form given later in footnote

[6] Equation (9) reveals the equality between the correspondent $h(q)$ minus the marginal cost of producing the final product and its raw material component.

[7] It should be stressed that the entire profit $\pi=[p_I-(c+k)]q_I$ is received by the monopolistic input supplier in Case (i) whereas the distribution of total profits in Case (iii) must be negotiated by the colluding firms, the lower limit for firm $B$ being set by $[p_I-(c+k)]q_{II}$.

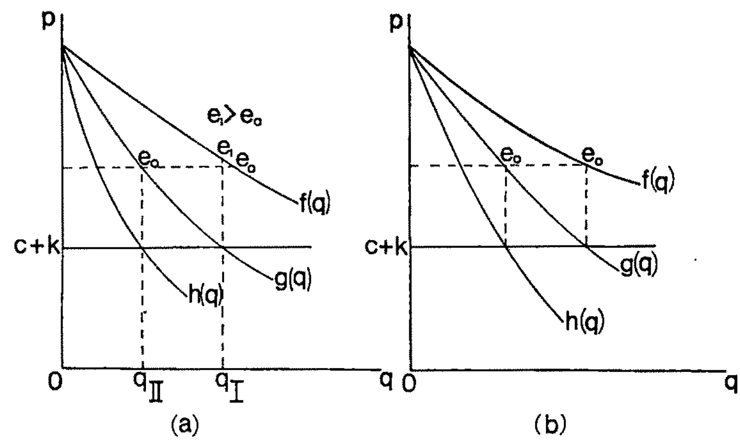



FIGURE 2

10; and 2) the elasticity of demand decreases with price, i.e., $de_f/dp>0$.[8]

To see the implication of proviso 1), consider arbitrarily two quantities $q_I$ and $q_{II}$ such that $f(q_I)=g(q_{II})$. It readily follows under proviso 1) that $e_f(q_I)=e_g(q_{II})$.[9] Combined with proviso 2), the present condition can be extended to the relations:

$$(10) \qquad e_f(q_I) \gtreqless e_g(q_{II})$$

$$\text{if and only if} \quad f(q_I) \gtreqless g(q_{II})$$

We are, accordingly, now in position to prove the generality of our first theorem.

[8] Neither of these specifications is required when $f$ is assumed to be of the specific algebraic form given in fn. 10. It warrants mention, however, in connection with the second constraint, that demand curves of constant or increasing elasticity are typically ruled out in the space economy since they generate the unlikely practice of discriminating against distant buyers, and this opens up the possibility of repackaging and resale by nearer buyers. For example, see Edgar Hoover, Arthur Smithies, and Greenhut, M. J. Hwang and Ohta. From the standpoint of spatial economics, therefore, the subject constraint is quite meaningful. Nevertheless, in fn. 10 below we shall include instances of increasing and constant elasticity that yield the same results.

[9] For proof consider the condition of $f(q_I)=g(q_{II})$. Then:

$$\frac{df}{dq_I}\frac{q_I}{f} = \frac{dg}{dq_{II}}\frac{q_{II}}{g}\frac{dq_{II}}{dq_I}\frac{q_I}{q_{II}}$$

$$\therefore\ e_f(q_I) = e_g(q_{II}) \quad \text{if } \frac{dq_{II}}{dq_I}\frac{q_I}{q_{II}} = 1 \text{ or } \frac{q_I}{q_{II}} = \text{constant}$$

To recall, Theorem I holds that the two alternative equilibrium conditions (6) and (9), i.e., $g(q_I)=c+k$ and $h(q_{II})=c+k$, generally imply $f(q_I)=g(q_{II})$, under well-defined demand functions given by the conditions specified above. Consider in this context the null hypothesis that $h(q_{II})=g(q_I)$ implies $f(q_I) \gtrless g(q_{II})$; via (10), this relation in turn implies $e_f \gtrless e_g$, as illustrated in Figure 2a. But this condition is impossible since combining (6) and (9) establishes:

$$(11) \quad f(q)\left(1-\frac{1}{e_f}\right) = g(q)\left(1-\frac{1}{e_g}\right), \qquad e_f,\ e_g > 1$$

$$\therefore\ e_f(q) \lesseqgtr e_g(q)$$

$$\text{if and only if} \qquad f(q) \gtreqless g(q)$$

In particular, note that $f(q) \gtrless g(q)$ implies $e_f \lesseqgtr e_g$, which contradicts the elasticity requirement derived from the null hypothesis and must therefore be rejected. The relations $g(q)=h(q)=c+k$ necessarily imply the remaining condition $f(q)=g(q)$, which condition alone consistently yields $e_f=e_g$ in (10) and (11). *Except for* the equality parts of the relations in (10) and (11), which alone are consistent, these equations reveal contradictory conditions derived from the equilibrium conditions

$g(q)=h(q)=c+k$. Under the defined demand function $f$ subject to the aforementioned constraints, the equilibrium conditions $g(q_I)=h(q_{II})=c+k$ yield not only $f(q_I)=g(q_{II})$ but also implies $e_g(q_I)=e_f(q_{II})$, as illustrated in Figure 2b.[10]

The second theorem can be readily established with the aid of the first theorem. Observe from the first theorem that the value of the equilibrium marginal revenue $g(q_{II})$ of the monopolist firm $A$ is the same under Case (ii) conditions as the price $f(q_I)$ set by the colluding monopolists of Case (iii). Since the price $f(q_{II})$ is in general higher than the corresponding marginal revenue $g(q_{II})$, it follows that $f(q_{II}) > f(q_I)=g(q_{II})$. This relation establishes formally the proposition that Case (iii) and Case (i) as well provide a lower equilibrium price than does Case (ii). Since $f$ is assumed to be a monotonically decreasing function of $q$, it further follows that $q_I > q_{II}$; i.e., the output produced under Cases (i) or (iii) is greater than that under Case (ii).

[10] A proof which utilizes the algebraic function $f(q)=a-bq^\alpha$ with varying combinations of the parameters $a$, $b$, and $\alpha$ is set forth below. Significantly, the combinations allowed include $a, b, \alpha>0$; $a, b<0$, $-1<\alpha<0$ (decreasing elasticity situations); and, in addition, $a=0$, $b<0$, $-1<\alpha<0$ as well as $a>0$, $b<0$, $-1<\alpha<0$, combinations which respectively establish economically stable equilibrium situations under conditions of constant and increasing demand elasticities as price decreases. Pursuant to the constraints set forth above on the function $f(q)=a-bq^\alpha$, the correspondents $g(q)$ and $h(q)$ are given by:

$$\text{(a)}\qquad g(q) = a - b(\alpha+1)q^\alpha$$

$$\text{(b)}\qquad h(q) = a - b(\alpha+1)^2q^\alpha$$

Let $g(q_I)=h(q_{II})=c+k$ to obtain the specific values $q_I$ and $q_{II}$:

$$\text{(c)}\qquad q_I = \left(\frac{a-c-k}{b(\alpha+1)}\right)^{1/\alpha}$$

$$\text{(d)}\qquad q_{II} = \left(\frac{a-c-k}{b(\alpha+1)^2}\right)^{1/\alpha}$$

Substituting $q_I$ of (c) in $q$ of the demand function $f(q)$ and $q_{II}$ of (d) in $q$ of the correspondent $g(q)$ yields:

$$\text{(e)}\qquad f(q_I) = a - \frac{a-c-k}{\alpha+1} = g(q_{II})$$

## II

The thesis presented in Section I applied to a perfectly competitive or perfectly monopolized final good market. However, what about relationships where the final good market is neither perfectly competitive nor perfectly monopolized? To evaluate this possibility, assume an oligopolistically competitive market of the Cournot type where each firm considers its rival's supply to be fixed. Profit maximization for the representative firm requires:

$$\text{(12)}\qquad \frac{\partial(pq_i)}{\partial q_i} = f(q) + f'\frac{\partial q}{\partial q_i}q_i = f(q) + f'(q)q_i = c_i + r_i \qquad i = 1, 2, \ldots, m$$

where $q_i$ stands for the $i$th firm's supply, $c_i$ for the $i$th firm's marginal production cost in using original input factors, and $r_i$ for the marginal cost of using the raw material input; alternatively, $r_i$ provides (in terms of $q$) the price charged the $i$th firm by the monopolistic input supplier. Note that the unit value applies to $\partial q/\partial q_i$ in (12) because market demand equals the aggregate supply of the $m$ oligopolistic firms, i.e.,

$$\text{(13)}\qquad q = \sum^m q_i$$

and any change in $q_i$ is assumed under Cournot's framework not to elicit a change in supply by rival firms. Summing both sides of (12) over all $i$, dividing by $m$, and utilizing (13) establishes:

$$\text{(14)}\qquad f(q) + \frac{f'(q)}{m}q = c + r$$

where $c = \dfrac{1}{m}\sum c_i \qquad r = \dfrac{1}{m}\sum r_i$

For the purpose of this paper, $c_i$ and $r_i$ must be assumed to be the same as they would be under conditions of perfect competition or monopoly, i.e., the same as $c$

and $r$ previously. Equation (14) may then be rewritten as:

$$(14') \qquad r = f(q) + \frac{f'(q)q}{m} - c$$
$$= f(q)\left(1 - \frac{1}{me_f}\right) - c$$

Equation (14′) represents the monopolistic input supplier's average revenue function in terms of $q$ under the specified Cournot conditions. When $m=1$, (14′) is equivalent to (8), the case of monopoly in the final good market. As $m$ increases, i.e., as more and more firms enter the final good market, the adjusted elasticity of demand $e_f^* = me_f$ increases by the multiplier $m$. And when $m$ approaches $\infty$, (14′) approaches (4) in reflection of perfect competition.[11]

To obtain the equilibrium price $r$ per unit of raw material input for all in-between market conditions (i.e., where $m$ is a finite number greater than one), some concrete form of the demand function $f$ must be assumed, such as the $f(q) = a - bq^\alpha$ evaluated previously in footnote 10. That function represents a general form as it generates not only convex, linear, or concave demands, but curves of either constant, decreasing, or increasing elasticities of demand depending on the alternative combinations of parameters $a$, $b$, and $\alpha$ that are applied. (See fn. 10.) Equation (14′) can then be particularized to:

$$(14'') \quad r = (a - c) - \left(\frac{m + \alpha}{m}\right) bq^\alpha$$

The profit maximizing condition for the monopolist firm $B$ is:

$$(15) \quad (a - c) - \left(\frac{m + \alpha}{m}\right) b(\alpha + 1)q^\alpha = k$$

where the left-hand side of (15) represents the raw material supplier's marginal revenue while the right-hand side, i.e., $k$, stands for his marginal cost of producing the raw materials. Substituting in (14″) the equilibrium $q$ obtainable from (15) yields:

$$(16) \qquad r = \frac{\alpha(a - c) + k}{\alpha + 1}$$

Remarkably enough, $r$ does not depend on the number of producers in the final good market; viz., the raw materials price $r$ charged the final good producer by the monopolistic input supplier is the same for any final good market type, whether it be simple monopoly (when $m=1$), duopoly ($m=2$), etc.[12] An invariance theorem on raw material inputs prevails regardless of selling conditions in the final good market under the specified Cournot behavioral assumption and the technical assumption on the form of the demand function.

A two-fold question remains with respect to oligopolistic competition in the final good market: 1) would horizontal merger between final good producers lower the market price of their good; and 2) would vertical merger between final good producers and the monopolistic supplier of raw materials lower the price of the final good? Answer to question 1) simply requires conceiving of a constant raw material price $r$ (in addition to $c$). Differentiating (14) or (14′) with respect to the number of competitive producers of the final good, yields—in conformance with (16)—the equation:

$$(17) \quad \frac{dp}{dm} = \frac{-(r + c)\left(e_f + m\dfrac{de_f}{dp}\dfrac{dp}{dm}\right)}{(me_f - 1)^2}$$

This equation demonstrates that under the

[11] See John Greenhut and M. L. Greenhut for a development along this line which isolates the impacts on delivered price schedules of the location of rivals and the intensity of their competition at alternative points.

[12] Manifestly we are not in a bilateral monopoly situation since our final good producer is not a monopsonist but simply an oligopolist or, as in Section I of the paper, a monopolist or perfect competitor.

normality assumption of decreasing elasticity with price, the equilibrium price $p$ must fall as the number of firms increases. It follows, accordingly, that horizontal merger, i.e., a decrease in $m$, raises the equilibrium commodity price $p$.

An asymmetric result obtains with respect to the vertical rationalization proposed under question 2). More specifically, profit maximization for any and all independent final goods producers is given by (12). But if these oligopolists, or just some of them, merge with the monopolistic supplier of the needed raw materials, the new profit maximization requires:

$$f(q) + f'(q)q_i = c_i + k_i, \quad i = 1, 2, \ldots, j \leq m \tag{18}$$

where (18) assumes that $(m-j)$ firms do *not* merge with the monopolistic supplier of raw materials, and hence equation (12) applies to these $(m-j)$ firms. Of course, the marginal cost $k_i$ of producing the raw material is lower than the comparable marginal cost (price) $r_i$ in (12). Summing (18) and (12) over their relevant ranges then gives:

$$mf(q) + f'(q)q = \sum_{i=1}^{m} c_i + \sum_{i=1}^{j} k_i + \sum_{i=j+1}^{m} r_i \tag{19}$$

where $$q = \sum_{i=1}^{j} q_i + \sum_{i=j+1}^{m} q_i$$

Dividing both sides of (19) by $m$ establishes:

$$f(q)\left(1 - \frac{1}{me_f}\right) = c + \frac{j}{m}k + \frac{m-j}{m}r = c + r - (r-k)j/m < c + r \tag{19'}$$

provided that $r > k$, i.e., the average price charged per package of raw materials by the monopolist supplier is greater than the average marginal cost of supplying that package. Thus the *average* marginal costs, viz.,

$$\left(\sum^{m} c_i + \sum k_i + \sum^{m} r_{j+i}\right) \Big/ m$$

decrease as more and more firms collude (merge) with the firm that supplies the raw materials. Given the market demand function $f(q)$ and the fixed total number $m$ of firms producing the final good, the same adjusted $MR$ curve $f(q)[1-1/me_f]$ that was derived from the original $MR = f(q) \cdot (1-1/e_f)$ obtains. Thus equilibrium price $f(q)$ *falls* as average $MC$ falls. Vertical merger lowers commodity price and increases the quantity demanded and the output produced.

## III

We have thus far assumed fixed proportions of inputs to output. The basic welfare results scored above remain unaffected, however, by variation in marginal costs. In demonstrating this condition, two concepts of variable marginal costs are required, as we shall extend our previous analysis to include spatial price relations.

Consider a monopolistic producer who ships his commodity to spatially separated market points. Let the demand function (1) apply now to any one local market point. If a transportation service is provided by a monopolistic carrier, the carrier's average revenue function would be given by (8), where $r$ would represent the freight rate applicable to the market point under consideration. The equilibrium condition for an independent monopoly situation between the carrier and shipper of goods is then given by (9), where $c$ now represents the marginal cost of production while $k$ provides the marginal cost of transportation. Note that $k$ and $c$ are specified differently in:

$$k = \phi(q) \tag{20}$$

$$c = \psi(Q), \qquad Q = \sum q \tag{21}$$

where the summation applies to all relevant market points. Thus $k$ is a direct function of $q$, but $c$ is not so related since a change in the particular supply $q_1$ may be carried out by an offsetting change in the supply $q_2$; the consequence would be that the total output produced $Q$ and the related cost $c$ are unchanged. For this reason, marginal production cost $c$ does not vary with a change in supply $q$ to a particular market point. We assume, however, that $c$ does increase if (and only if) *aggregate* $q$ (i.e., $Q$) increases.

The relation between $c$ and $k$ with respect to a change in $q$ is illustrated in Figure 3. Marginal cost $k$ of shipping goods is assumed in the figure to be a decreasing function of $q$; i.e., economies derive from using larger input quantities, with the effect $c+k$ decreases as $q$ increases.[13] Suppose equilibrium prevails initially at $D$ in a particular market, and that the input supplier (carrier) and his customer are independent monopolists. The question of impacts of collusion may, accordingly, be raised again.

Any change in equilibrium from, say, $D$ to $D'$ implies an increase in a particular market's supply as price is lowered. Collusion must, however, change the equilibrium price-quantity set *for all submarkets*. The consequence is that total output and cost of production increase (compare $c$ with $c'$ in Figure 3); correspondingly, the $MC$ curve shifts from $c+k$ to $c'+k$, with final equilibrium taking place at point $D''$. This new position guarantees a greater output than that obtainable when the monopolists are independent. But the question arises as to whether or not an increase in $c$ which results from collusion might occasionally be so great as to generate a smaller equilibrium output. To resolve this particular question, note initially that $c$ rises only if *equilibrium* output increases; in

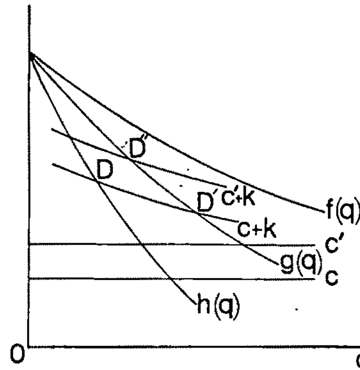



FIGURE 3

other words, marginal production cost $c$ increases only if the *equilibrium* output produced under Case (iii) conditions increases. If the equilibrium output does not increase, neither does the related marginal production cost. It further follows that the rightward placement of $g(q)$ compared to $h(q)$ cannot be offset by a change in $c$ to $c'$, and accordingly output must increase. This requirement fully establishes the generality of our proposition that output is greater and price lower for Case (iii) than Case (ii).

## IV

The comparative profitability of merged firms with that of individual monopolists is thus completely given by the demand condition $f(q)$. Maximum profits obtain if and only if marginal revenue $g(q)$ is equated with the lowest obtainable marginal cost, i.e., $c+k$ distinct from $c+r$. Any deviation from this equilibrium, such as in Case (ii), necessarily implies a decrease in profits.

The results obtained above apply to any particular market in which sales are made. Aggregate outputs and profits increase and prices are lower when vertical merger or some other rationalization process takes place between an input supplier and the user(s) of his product whenever the succes-

[13] Decreasing marginal cost $k$ is not a required assumption for our proof.

sive stages of production involve independent monopolistic powers. The surprising implication of these relationships is that excess capacity may be eliminated to a considerable extent by rationalization in related industries. Moreover, greater efficiency in production derives from the integration process described in this paper under imperfect market conditions.

The last contention is a sweeping one which is underscored by the fact that production generally requires the input of raw materials, intermediate goods, and services. Given a demand function for a good, one may deduce a derived demand function for, let us say, the intermediate good or service, such as transportation. This function would yield, in turn, another derived demand function, for example, for the tires used by the carrier, and so forth. Successively shrinking *correspondents* to derived demand functions thus obtain, as the smallest correspondent is equated with the relevant marginal cost to provide the equilibrium applicable under the case of independent monopolistic firms. This smallest correspondent obviously shrinks toward the origin as more and more related industries are subjected to independent monopoly powers. And the converse principle holds, as the smallest *correspondent* would be twisted upward by integration of firms in related industries. Equilibrium output increases as a result, except for the unlikely *and unstable* cost situation of rapidly decreasing $MC$ in the region of the equality of marginal values that prevailed before the integration of firms.

In a related context to this paper, Donald Dewey contended that rationalization of firms increases efficiency in imperfectly competitive markets. This effect was said to arise since the merged firms can utilize one plant more efficiently in producing a small high-cost output than can two firms which would otherwise produce at very inefficient small output points. Only over a narrow range of outputs, where each plant must be utilized, would the costs of the unified firm necessarily involve the sum of the costs of *each* of the individual firms. In fact, Dewey argued that *even in such case* the unified firm would operate the single firm's plant(s) more efficiently at greater profit than that which would eventuate under a Cournot competitive equilibrium. The Dewey merger model unfortunately involves smaller output than the competitive model, and in this respect can be criticized by welfare theorists.

Our paper, similar to Dewey's, has *also* advocated merger as a source of increased efficiency in imperfectly competitive markets. But even closer to Wu's, it proposes vertical rationalization only for firms in related industries. We contend that merger of firms in related industries generates a greater output than the sum of the individual outputs of nonintegrated firms. In effect, this present paper has, therefore, illustrated *and* proved two theorems which were *implicitly* assumed by the present authors and Hwang in this *Review* on the type of (and direction of) discrimination practiced by motor carriers. That paper, in particular, anticipated the advantages to carriers and shippers of vertical integration in noting that freight rates were unaffected by the market structure in the shipper's industry (compare Josephine Olson). The roots of the present paper based as they are on successive stages of production thus include the distribution process as well as the production process. Indeed, one can readily conceive of freight rates covering certain specified places that converge in time under Commission regulations to the derived $r_i$ values that were described above. By permitting shippers of goods the ownership of their own carriers (or permitting them vertical integration with existing contract or common carriers), cost $k_i$ replaces $r_i$ and in the process generates greater outputs and lower prices.

Other policy implications of our theorems should be manifest. For example, objections being expressed against the possible "take-over" of the retailing function by the larger oil companies under deregulation of "old oil prices" appear ill-founded as greater output and lower prices should result from that integration. Excluded from our analysis of related markets is only the possibility of bilateral monopoly, since our higher stage producer is not conceived to have monopsony buying power over the lower stage producer. All other related market situations fall within our analysis. Possibly most important of all are the antitrust implications of our findings which, to say the least, will prove shocking to many.[14] At the same moment, these implications will clearly warrant study.

[14] Consider a standard case before our courts involving vertical merger, such as United States v. Yellow Cab. Concern in that case was with a monopolistic buying firm (e.g. Yellow Cab or our firm $A$) which would exclude other input suppliers from certain markets if it merged vertically with a particular supplier of the needed input. In opposite situations, elimination of competition on the final good market level might appear questionable to many. However, strict interpretation of our theorems must yield an interesting difference in conclusions. A less rigid interpretation from a long-run perspective can be found in Greenhut, p. 366.

## REFERENCES

**W. J. Baumol and D. F. Bradford**, "Optimal Departures From Marginal Cost Pricing," *Amer. Econ. Rev.*, June 1970, *60*, 265–83.

**A. Cournot**, *Mathematical Principles of the Theory of Wealth*, translated by N. T. Bacon, New York 1927.

**D. Dewey**, *The Theory of Imperfect Competition: A Radical Reconstruction*, New York 1969.

**J. Greenhut and M. L. Greenhut**, "Spatial Price Discrimination, Competition, Locational Effects," *Economica*, Nov. 1975, *42*, 401–19.

**M. L. Greenhut**, *A Theory of the Firm in Economic Space*, Austin 1974.

———, **M. J. Hwang, and H. Ohta**, "Price Discrimination by Regulated Motor Carriers: Comment," *Amer. Econ. Rev.*, Sept. 1974, *64*, 780–84.

**E. M. Hoover**, "Spatial Price Discrimination," *Rev. Econ. Stud.*, June 1937, *4*, 182–91.

**F. Machlup and M. Taber**, "Bilateral Monopoly, Successive Monopoly, and Vertical Integration," *Economica*, May 1960, *27*, 101–19.

**J. Olson**, "Price Discrimination by Regulated Motor Carriers," *Amer. Econ. Rev.*, June 1972, *62*, 395–402.

**J. Robinson**, *The Economics of Imperfect Competition*, London 1933.

**A. Smithies**, "Monopolistic Price Policy in a Spatial Market," *Econometrica*, Jan. 1941, *9*, 63–73.

**S. Y. Wu**, "The Effects of Vertical Integration On Price and Output," *West. Econ. J.*, May 1964, *2*, 117–33.

**United States v. Yellow Cab**, 332 US 218, 1947.

# The Effect of Rate of Return Regulation is Highly Sensitive to the Nature of the Uncertainty

*By* YORAM C. PELES AND JEROME L. STEIN*

The fundamental conclusion derived from the theory of the monopoly firm seeking to maximize profits subject to a regulatory constraint is that: the optimum *ex ante* scale of plant chosen by the firm will increase as the regulated rate of return ($s$) declines towards the cost of capital ($i$). This proposition and its implications, known as the Averch-Johnson (A-J) effects,[1] have been the focus of numerous articles on rate of return regulation.

There is a serious danger that public policy concerning public utility regulation will be based upon an uncritical acceptance of the A-J effects. Although uncertainty of the state of nature characterizes the investment decisions, the A-J theorems have been based upon a model which ignores uncertainty. In the present paper we prove that the fundamental A-J conclusion is not robust to the type of uncertainty faced by the firm.

Decisions made by the firm are dichotomized into *ex ante* decisions (controls) made before the state of nature ($u$) is known and *ex post* decisions (controls) made after the state of nature is known. The scale of plant ($K$) is an *ex ante* control. The quantities of the variable inputs and the price charged by the firm are *ex post* controls. Uncertainty is introduced into the maximal quasi-rents function $R(K, u)$, which is defined as the total quasi rents that will be earned when the firm with capital stock $K$ selects the *optimum ex post* controls, and the state of nature is $u$. The risk-neutral firm selects the scale of plant to maximize expected profits $E[R(K, u)] - iK$, where $E$ is the expectations operator.

If the uncertainty concerning the state of nature is additive so that $R(K, u) = R(K)+u$, and $u$ is independent of scale $K$, the conventional A-J results are obtained. The imposition of an effective regulatory constraint will either drive the firm out of business if the spread $s-i$ is not sufficiently large, or it will lead the firm to select *ex ante* a scale of plant larger than that chosen by the unregulated monopolist. When the uncertainty is multiplicative, $R(K, u) = R(K)(1+u)$, the conventional A-J effect is completely reversed. Assume that the regulatory constraint does not drive the firm out of business. Then the closer the regulated rate of return is to the cost of capital (but $s>i$), the *smaller* will be the optimum *ex ante* scale of plant relative to that chosen by the unregulated monopolist. Call this the "Anti A-J Theorem."

The effects of rate of return regulation are highly sensitive to the characterization of the uncertainty. For this reason, public policy towards public utilities should not be formulated on the basis of the A-J effect until there is evidence that the uncertainty is not of the multiplicative type. Otherwise, tightening the regulatory con-

* The Hebrew University, University of Florida; and Brown University, respectively. We thank F. Arditti, W. Baumol, G. Borts, H. Levy, R. Lusky, and E. Sheshinski for their valuable criticisms of an earlier draft.

[1] See William Baumol and Alvin Klevorick, Stein and George Borts for a discussion of these issues and a bibliography of the salient articles in the area.

straint may lead to a decrease in output as well as to production inefficiency. No regulation would then be socially preferable to this sort of regulation.

## I. The Scale of Plant is the *Ex Ante* Control and the Variable Inputs are the *Ex Post* Controls

### A. *The Maximal Quasi-Rents Function*

When the firm decides upon its optimal scale of plant, the total revenue function expected to prevail when the plant is in operation is not known with certainty. A firm does not know precisely what will be the growth of demand, the geographic location of its customers, or the development of rival technologies by firms which produce substitute products. If the proposed plant scale involves the introduction of a new technology, then the firm is not even sure what will be the rate of output associated with a given vector of inputs. Uncertainty in this context refers to the state of nature, described by variable $u$, that prevails when the plant is in operation.

A firm which seeks to maximize profits from a given plant when the state of nature is revealed selects the quantities of variable inputs—referred to under the generic name of labor—to maximize *quasi rents* $Z$. The labor input is an *ex post* control.

Quasi rents $Z$ earned by a firm are total revenue $pQ$ (where $p$ is price and $Q$ is quantity sold) less total variable costs $WL$, where $L$ is the quantity of the labor input and $W$ is the exogenously given price of labor. Price $p$ depends upon $Q$ and the state of nature: $p=f(Q, u)$. Quantity sold is assumed to equal quantity produced, where the latter depends upon the inputs of labor and capital (i.e., scale of plant) $K$. For these reasons, total revenue can be written as a function $H(K, L; u)$ of the inputs and the state of nature. Function $H$ is strictly concave in the inputs and reaches a maximum. Quasi rents are then described by equation (1).

$$Z = H(K, L; u) - WL \tag{1}$$

*Ex post* control $L$ is chosen to maximize quasi rents $Z$, given *ex ante* control $K$ and the state of nature $u$. Under these conditions, the marginal revenue product of labor $H_L(K, L; u)$ is set equal to the nominal wage $W$ as described by equation (2).

$$H_L(K, L; u) = W \tag{2}$$

Solve equation (2) for the optimal quantity of labor and derive equation (3). The optimum quantity of labor (*ex post* control) is a function of the scale of plant (*ex ante* control), the nominal wage, and the revealed state of nature. Since the nominal wage is a constant, it will not always be written explicitly. We assume that an increase in $K$ always raises the marginal revenue product of labor in the relevant range: $H_{LK}>0$.

$$L = l(K; u, W); \qquad l_K = \frac{H_{LK}}{-H_{LL}} > 0, \tag{3}$$
$$l_u = \frac{H_{Lu}}{-H_{LL}} > 0, \quad l_W = \frac{1}{H_{LL}} < 0$$

Substitute the optimal quantity of labor into equation (1) and derive equation (4). The latter represents the *maximum* quasi rents that can be obtained by a firm with a scale of plant $K$ when the state of nature is $u$.

$$\begin{aligned} R &= H[K, l(K; u, W); u] \\ &\quad - Wl(K; u, W) \\ &= R(K; u, W) = R(K, u) \end{aligned} \tag{4}$$

Again, remember that $R$ is also a function of $W$. The $R(K, u)$ function is graphed in Figure 1 as $OFMBR$. The slope of this function is the marginal revenue product of capital:

$$R_K(K, u) = H_K(K, L; u) \tag{4'}$$

and it is strictly concave because

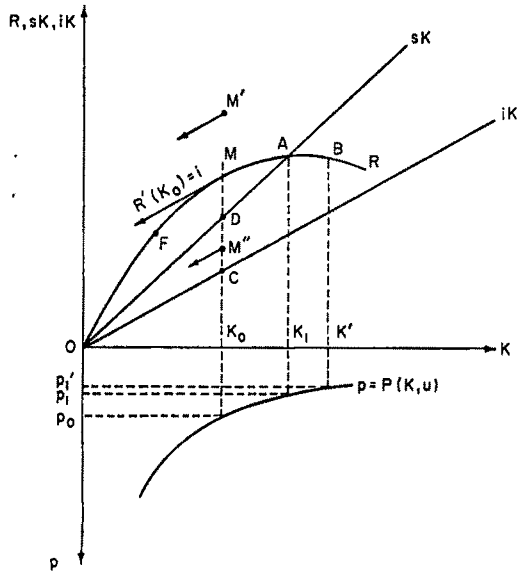



FIGURE 1. CURVE $OMR$ REPRESENTS MAXIMAL QUASI RENTS. RAY $OsK$ REPRESENTS TOTAL PERMISSIBLE PROFITS AT RATE OF RETURN $s$. RAY $OiK$ REPRESENTS TOTAL CAPITAL COSTS. CURVE $p=p(K, u)$ DETERMINES THE OPTIMAL PRICE. THE UNREGULATED MONOPOLIST EMPLOYS CAPITAL STOCK $K_0$. THE A-J POINT IS $K_1$.

$$(4'') \quad R_{KK}(K, u) = H_{KK}(K, L; u) - \frac{H_{KL}^2}{H_{LL}}(\ ) < 0$$

The optimal price $p$ charged by the firm is described by equation (5), which is drawn in Figure 1 (lower part). It corresponds to the point on the demand curve above the intersection of the marginal revenue curve (associated with state of nature $u$) with the short-run marginal cost curve (associated with scale of plant $K$).

$$(5) \quad p = P(K; u, W); \qquad P_K(\ ) < 0$$

Function $P_K$ is negative because a rise in the scale of plant shifts the short-run marginal cost curve to the right, and it intersects the marginal revenue curve at a higher level of output. Since the demand curve is negatively sloped, the price declines as the optimum output increases.

## B. *Alternative Characterizations of the Uncertainty*

There is a considerable literature concerning the theory of the firm facing an uncertain demand. One conclusion is that: if the firm's *ex ante* control is quantity produced and offered for sale, then the risk-neutral firm would select the same optimum quantity as it would if it were certain of the state of nature.[2] In our analysis, the scale of plant is the *ex ante* control, and the resulting state of nature determines the total revenue function. To stay as close to A-J as we can, assume that the firm is risk neutral.[3] Uncertainty would not affect the *ex ante* scale of plant selected by the *unregulated* firm. Nevertheless, we prove that uncertainty concerning the state of nature may induce the *regulated* firm to select a smaller scale of plant than would be chosen if there were no uncertainty concerning the state of nature.

There are several ways to characterize the uncertainty concerning the state of nature that will prevail when the plant is in operation. Our main theme is that the nature of the uncertainty will profoundly affect the response of the firm to a rate of return constraint.

In equation (6), uncertainty is characterized as a percentage deviation of the maximal quasi-rents function $R(K, u)$ from its expected value $E[R(K, u)]$. The coefficient of variation, defined as the ratio of the standard deviation to the mean, is a constant independent of $K$.

$$(6) \quad R(K, u) = R(K)(1 + u)$$

$$(6') \quad E[u] = 0$$

$$(6'') \quad E[u^2] = \int u^2 f(u)du > 0$$

$$(6''') \quad \underset{u}{E}[R(K, u)] = R(K)$$

[2] See Agnar Sandmo, and Hayne Leland, p. 283.
[3] The assumption of risk aversion would bias the results away from Averch-Johnson.

Multiplicative deviation $u$ is a random variable with a zero mean, positive variance, and a probability density function $f(u)$. Consequently, the expected value of the maximal quasi rents is $R(K)$.

Alternatively, the uncertainty could be additive and independent of $K$ as described by equation (7). Again, let the expected value of the disturbance be equal to zero with a positive variance and a probability density function $g(u)$. The coefficient of variation in this case is a declining function of $R$.

$$R(K, u) = R(K) + u \tag{7}$$

$$E[u] = 0 \tag{7'}$$

$$E[u^2] = \int u^2 g(u) du > 0 \tag{7''}$$

$$\underset{u}{E}[R(K, u)] = R(K) \tag{7'''}$$

If the *unregulated* firm were risk neutral, the *ex ante* control $K$ would be chosen to maximize expected profits $\pi_0(K)$, the difference between the expected maximal quasi rents $E[R(K, u)]$ and total capital costs $iK$, where $i$ is the cost of capital. The zero subscript denotes the absence of regulation.

$$\begin{aligned} \underset{K}{\text{Max}}\, \pi_0(K) &= \underset{K}{\text{Max}}\, \underset{u}{E}[R(K, u)] - iK \\ &= \underset{K}{\text{Max}}\, R(K) - iK \end{aligned} \tag{8}$$

Since the expected quasi rents $E[R(K, u)]$ equal $R(K)$ in both the additive and the multiplicative cases, the *ex ante* scale of plant chosen will be the same as would be chosen if there were no uncertainty. The expected marginal product of capital $(d/dK)E[R(K, u)] = E(d/dK)R(K, u) = R'(K)$ in both the additive and the multiplicative cases. An unregulated risk-neutral firm will select $K_0$ (in Figure 1) as the optimum *ex ante* control, where the expected marginal revenue product of capital $R'(K_0)$ is equal to the cost of capital $i$, provided that $\pi_0(K_0) \geqq 0$.

$$R'(K_0) = i \tag{9}$$

This point is described by point $M$ in Figure 1. Total profits are maximal and positive at point $M$, and the monopolist will charge price $p_0$. The ratio of quasi rents to the optimal capital stock is $K_0M/OK_0$. The nature of the uncertainty is irrelevant for the risk-neutral unregulated monopolist.

## II. The Regulatory Constraint and the Expected Maximal Quasi-Rents Function

When a regulatory constraint is imposed the firm is not allowed to earn more than $sK$ of quasi rents. To be effective, the regulated rate of return $s$ must be less than what was earned by the unregulated monopolist, but it must exceed the cost of capital $i$. In terms of Figure 1, $K_0M/OK_0 > s > i$.

The expected maximal quasi rents consists of two parts. Define $S$ as the set of states of nature $u$ where the maximal quasi rents $R(K, u)$ are less than $sK$. Define $T$ as the complement of that set.

$$S \equiv [u \mid R(K, u) < sK] \tag{10a}$$

$$T \equiv [u \mid R(K, u) \geqq sK] \tag{10b}$$

$$Pr[u \in S] + Pr[u \in T] = 1 \tag{10c}$$

Suppose that the firm selected a scale of plant $K'$ and the state of nature that occurred when the plant came into operation yielded maximal quasi-rents curve $OFMBR$ in Figure 1. The state of nature is in set $S$ because maximal quasi rents $K'B$ are less than $sK'$. At that scale of plant and given the state of nature, profits would be maximized by charging price $p'$. Nevertheless, the ratio of quasi rents to the stock of capital would be below $s$, the regulated rate. There is nothing that the regulatory authorities can do to help the firm earn $sK'$ because $K'B < sK'$ represents the maximal quasi rents that can be generated from this plant. By raising the price above $p'$, the firm would be generating less than the

maximum profits associated with this scale of plant and state of nature: i.e., it would be operating below point $B$ in Figure 1.

Had the firm selected scale of plant $K_0$, when the state of nature is in set $T$ yielding maximal quasi rents $K_0M$, then it would be earning a rate of return in excess of $s$. The maximum profit price charged by the firm would be $p_0$. During the period of the regulatory lag, the firm would be earning a rate of return greater than the "fair rate." Sooner or later, the regulatory authority would catch up with the firm and require it to lower its price below $p_0$.[4] At the lower price, the quantity demanded would rise. The regulated monopolist is required to meet the demand. To produce the greater quantity demanded, more variable inputs (labor) would have to be employed with scale of plant $K_0$. At point $M$ on the maximal quasi-rents curve, the marginal revenue product of labor is equal to the wage: equation (2). When employment increases to satisfy the greater quantity demanded at the lower price, the marginal revenue product of labor is driven down below the wage. The regulatory authority will require the firm to lower its price and meet the quantity demanded until total quasi rents fall to $K_0D=sK_0$, which are below the maximal quasi rents $K_0M$.

The expected quasi rents, denoted by $E[R]$, are described by equation (11). When state of nature $u$ is in set $S$, then the firm earns the maximal quasi rents $R(K, u) < sK$. When the state of nature is in set $T$ then, after the regulatory lag, the firm earns $sK \leqq R(K, u)$ of quasi rents. The second integral in equation (11) is the probability that state of nature $u$ is in set $T$.

$$(11) \quad E[R] = \int_{u\in S} R(K, u)f(u)du + sK\int_{u\in T} f(u)du \leqq sK$$

Suppose that set $S$ is not empty. Then for $u$ in set $S$, the maximal quasi rents $R(K, u)$ are less than $sK$.

$$(11') \quad E[R] = \int_{u\in S} R(K, u)f(u)du + sK\, Pr[u \in T] < sK$$

It follows that the expected quasi rents earned by a regulated monopolist are less than $sK$, the maximum permissible quasi rents.[5]

## III. Uncertainty and the Regulated Monopolist

### A. *When the Uncertainty is Additive, Conventional Results are Obtained*

If the disturbance term is additive and independent of $K$, the conventional A-J results are obtained. The imposition of an effective regulatory constraint will either drive the firm out of business or will lead it to select *ex ante* a scale of plant larger than $K_0$ chosen by the unregulated monopolist.

Equation (12) describes the expected quasi rents earned by a regulated firm when the disturbance term is additive. The first integral in equation (12) describes the maximal quasi rents when $R(K)+u$ is less than $sK$, i.e., when $u$ is in set $S$. The second integral in equation (12) describes the situation when the quasi rents are at least as great as $sK$, i.e., $u$ is in set $T$.

$$(12) \quad E[R(K) + u] = \int_{u\in S} [R(K) + u]g(u)du + sK\int_{u\in T} g(u)du$$

Equation (12′) is a useful way of writing (12). The integral $\int_{u\in S} ug(u)du$ is the expected value of $u$ such that $R(K)+u$ is

[4] It is assumed that the regulatory lag is short relative to the life of the plant. For this reason, the regulatory lag is disregarded here.

[5] This is the essence of corollary 2 in the paper by Stylianos Perrakis, appearing elsewhere in this issue.

less than $sK$, and can be written as $E[u|u \in S]$.

$$(12') \quad E[R(K)+u] = R(K)\, Pr[u \in S] + \int_{u \in S} ug(u)du + sK\, Pr[u \in T]$$

Total expected profits denoted by $\pi_1(K)$ are expected quasi rents minus total capital costs $iK$, as described by equations (13) and (13′). Subscript 1 denotes the profits of the regulated firm when the uncertainty is additive.

$$(13) \quad \pi_1(K) = E_1[R(K)+u] - iK = R(K)\, Pr[u \in S] + E[u|u \in S] + sK\, Pr[u \in T] - iK$$

Since the state of nature must be either in set $S$ or set $T$, equation (13) may be written as:

$$(13') \quad \pi_1(K) = [R(K) - sK]\, Pr[u \in S] + (s-i)K + E[u|u \in S]$$

Now it is easy to prove that the optimal *ex ante* scale of plant chosen by the regulated monopolist will exceed $K_0$, the scale selected by the unregulated monopolist provided that the regulation does not drive the firm out of business completely.

Differentiate $\pi_1(K)$ with respect to $K$. Term $\phi_1(K)$ involves the derivatives of $Pr[u \in S]$ and $E[u|u \in S]$, and will generally be ignored.

$$(14) \quad \pi_1'(K) = [R'(K) - s]\, Pr[u \in S] + (s-i) + \phi_1(K)$$

The firm will go out of business if Max $\pi_1(K)$ is negative. Assume that $(s-i)$ is sufficiently large that the maximum expected profits are positive at the unregulated point $\pi_1(K_0) > 0$, so that this possibility will not occur.

Evaluate equation (14) at the unregulated point $K_0$ where $R'(K_0) = i$ and derive:

$$(14') \quad \pi_1'(K_0) = (s-i)\, Pr[u \in T] > 0$$

It follows that total expected profits can be increased by selecting *ex ante* a scale of plant larger than that chosen by the unregulated monopolist.

In the conventional A-J case when there is no uncertainty, it is certain that the firm will earn more than $sK_0$ when it selects scale $K_0$. Point $M$ is above ray $OsK$ in Figure 1 with probability 1; $Pr[u \in T] = 1$ at that point. Consequently, $\pi_1'(K_0) = (s-i) > 0$, and the firm selects a larger scale of plant than it did when it was not regulated. The well-known A-J effect may be considered as a special case of inequality (14′) above.

### B. *When the Uncertainty is Multiplicative, an Anti A-J Theorem is Derived*

When the uncertainty is multiplicative, the conventional A-J effect is completely reversed. The closer the regulated rate of return is to the cost of capital, the *smaller* will be the optimum *ex ante* scale of plant. Call this the anti A-J theorem.

Total expected profits of the regulated firm are $\pi_2(K)$, where subscript 2 denotes the regulated firm facing a multiplicative disturbance.

$$(15) \quad \pi_2(K) = E[R(K)(1+u)] - iK$$

Expected maximal quasi rents are described by equation (16). The first integral describes the situation where the maximal quasi rents $R(K)(1+u)$ are less than $sK$: i.e., the state of nature $u$ is in set $S$. The second integral refers to the situation where the firm is able to earn at least the regulated rate of return: i.e., $u$ is in set $T$, but would have to lower its price to satisfy the regulatory constraint.

$$(16) \quad E_2[R(K)(1+u)] = \int_{u \in S} R(K)(1+u)f(u)du +$$

$$sK \int_{u \in T} f(u)du < sK$$

Equation (16′) is another way of writing equation (16). Certainly for scales of plant between 0 and $K_1$ in Figure 1, the expected value of $u$ which will make $R(K)(1+u) < sK$ is negative. Integral $\int_{u \in S} uf(u)du \equiv E[u|u \in S]$ is denoted by $-\gamma$ in equation (16″).

$$(16') \quad E_2[R(K)(1+u)] = R(K)\,Pr[u \in S] + R(K)E[u \mid u \in S] + sK[1 - Pr[u \in S]]$$

$$(16'') \quad E[u \mid u \in S] = -\gamma < 0$$

Total profits expected by the regulated monopolist when the uncertainty is multiplicative are given by equation (17), derived from equations (15) and (16).

$$(17) \quad \pi_2(K) = (R(K) - sK)\,Pr[u \in S] + (s - i)K - \gamma R(K)$$

Differentiate equation (17) with respect to $K$. Term $\phi_2(K)$ involves the derivatives of $Pr[u \in S]$ and $\gamma$. As a rule, it will be ignored.

$$(18) \quad \pi_2'(K) = (R'(K) - s)\,Pr[u \in S] + (s - i) - \gamma R'(K) + \phi_2(K)$$

Evaluate this expression at the unregulated point $K_0$ where $R'(K_0) = i$. Equation (19) is obtained by setting $R'(K_0) = i$ and $Pr[u \in T] = 1 - Pr[u \in S]$.

$$(19) \quad \pi_2'(K_0) = (s - i)\,Pr[u \in T] - \gamma i$$

At scale of plant $K_0$, the expected value of the maximal quasi rents $R(K_0)$ exceeds $sK$. Point $M$ in Figure 1 is above ray $OsK$. Consequently at point $K_0$, the expected value of $u$ such that $u$ is in set $S$ is negative. Variable $\gamma$ is positive.

Equation (19) describes the slope of the expected profits function at the unregulated optimum scale of plant. As the regulated rate of return decreases, $\pi_2'(K_0)$ will become negative; and the optimum *ex ante* scale of plant selected by a regulated firm will be less than that chosen by the unregulated monopolist. This is an anti A-J theorem.

The conventional A-J effect can be considered as a special case of our analysis. When there is certainty that point $M$ lies above ray $OsK$ then $Pr[u \in T] = 1$ and $\gamma = 0$ at scale of plant $K_0$. Then $\pi_2'(K_0) = (s-i) > 0$ at the unregulated point. Total expected profits can be raised by selecting an *ex ante* scale of plant in excess of $K_0$, which is the A-J effect. Once multiplicative uncertainty is introduced, a sufficiently large gap $(s-i)/i > \gamma/Pr[u \in T]$ must exist between the regulated rate of return and the cost of capital to induce the firm to select *ex ante* a scale of plant in excess of the unregulated scale $K_0$.

### C. *Discussion and Interpretation of the Results*

Every theorem, once proved and understood, should be obvious. Our theorem was that the character of the uncertainty determines whether a decline in the regulated rate of return towards the cost of capital raises or lowers the optimum *ex ante* scale of plant relative to what would be chosen if there were no regulation. Conventional A-J results are derived when the uncertainty is additive and independent of the scale of plant in the maximal quasi rents function. If, however, the uncertainty is multiplicative in the maximal quasi-rents function, then a decline in the regulated rate of return towards the cost of capital will lower the optimum *ex ante* scale of plant relative to what would be chosen if there were no regulation. As far as we know, these points have not been recognized in the literature.

These results may not as yet be obvious, but they can better be understood on the basis of more specific examples. First, consider the additive case where the distur-

bance is independent of the scale. Let $R(K)$ be described by curve $OMR$ in Figure 1, where $R(K)$ exceeds $sK$ for values of $K$ in the open interval $(0, K_1)$. Let there be three possibilities.

| $R(K, u)$ | Probability |
|---|---|
| $sK < R(K) + u$ | $p/2$ |
| $sK = R(K)$ | $1 - p$ |
| $sK > R(K) - u$ | $p/2. = Pr[u \in S]$ |

Then the total expected profits function $\pi_1(K)$ can be written as equation (20) or (20′), and it is based upon equation (13).

$$\pi_1(K) = (1 - p/2)sK + [R(K) - u](p/2) - iK \tag{20}$$

or

$$\pi_1(K) = (1 - p/2)(s - i)K + [R(K) - u - iK](p/2) \tag{20'}$$

Maximal expected profits are a linear combination of $(s-i)K$ with probability $(1-p/2)$ and $[R(K)-u-iK]$ with probability $p/2$.

Marginal expected profits are:

$$\pi_1'(K) = [(1 - p/2)s + (p/2)R'(K)] - i \tag{21}$$

$$\pi_1'(K) = (1 - p/2)(s - i) + (p/2)[R'(K) - i] \tag{21'}$$

In terms of Figure 1, maximal quasi rents will be above curve $OMR$ with probability $p/2$, below that curve with probability $p/2$ and will be on the curve with probability $(1-p)$. If the maximal quasi-rents curve is $R(K)$ or $R(K)+u$, then the marginal quasi rents will be $sK$ due to regulation; and marginal profits will be $(s-i)$. If the maximal quasi-rents curve is $R(K)-u$, then the marginal rents will be $R'(K)$; and the marginal profits will be $R'(K)-i$. In the additive case, when $u$ is independent of the scale, the slope of the maximal quasi-rents curve is independent of the disturbance term. It just depends on the scale of plant. The expected marginal rents are:

$$[(1 - p/2)s + (p/2)R'(K)] \quad \text{independent of } u$$

At the unregulated optimum scale of plant $K_0$, the marginal quasi rents will be $s$ with probability $(1-p/2)$ if the state of nature puts the firm at point $M$ or $M'$ in Figure 1; and the marginal quasi rents will be $R'(K_0)=i$ with probability $p/2$ if the state of nature puts the firm at point $M''$. The weighted average of $s$ and $i$ exceeds $i$ at the unregulated point.

$$\pi_1'(K_0) = (1 - p/2)(s - i) = (s - i)\, Pr[u \in T] > 0 \tag{21''}$$

Since $\pi_1'(K_0)$ is positive, expected profits can be raised by selecting a scale of plant greater than $K_0$, the optimum *ex ante* scale of plant chosen by the unregulated monopolist. Uncertainty, however, may deter the firm from investing in this industry, unless there is a sufficiently large gap between $s$ and $i$. For example, suppose that the firm selected the A-J solution $K_1$ where $R(K_1)=sK_1$. Then total expected profits are:

$$\pi_1(K_1) = (s - i)K_1 - up/2 = (s - i)K_1 + E[u \mid u \in S] \tag{22}$$

In the certainty case, these profits will be positive. When there is uncertainty, the value of $(s-i)$ must be positively related to the riskiness of the venture if expected profits are to be positive. The greater $u(p/2)$, a reasonable measure of risk in this case, the greater must be the spread between the regulated rate of return and the cost of capital, in order to justify investment in this industry.

A very different situation prevails in the multiplicative case. Consider the following distribution of quasi rents. Suppose that $K_1 > K > 0$ in terms of Figure 1 below.

| $R(K, u)$ | Probability |
|---|---|
| $sK < R(K)(1+u)$ | $p/2$ |
| $sK = R(K)$ | $1-p$ |
| $sK > R(K)(1-u)$ | $p/2 = Pr[u \in S]$ |

When there is regulation, total expected profits are:

$$\pi_2(K) = sK(1-p/2) + R(K)(1-u)(p/2) - iK \tag{23}$$

Expected quasi rents will be $sK$ with probability $1-p/2$ due to the regulatory constraint and $R(K)(1-u)$ with probability $p/2$ in the unfavorable case. Marginal expected profits will be

$$\begin{aligned}\pi_1'(K) &= [s(1-p/2) + R'(K)(1-u)(p/2)] - i \\ &= (s-i)(1-p/2) + [R'(K)(1-u) - i](p/2)\end{aligned} \tag{24}$$

If the total maximal quasi-rents curve is above $sK$, then the marginal quasi rents will be $s$ with probability $(1-p/2)$. If the maximal quasi-rents curve is below $sK$, then the marginal quasi rents will be $R'(K)(1-u)$ with probability $p/2$. The greater the range of outcomes, the larger is $u$; consequently, the smaller will be the marginal returns $R'(K)(1-u)$ in the case when the maximal quasi-rents curve is below $sK$. Since the expected marginal rents are a weighted average of $s$ and $R'(K)(1-u)$, $\pi_2'(K)$ is negatively related to $u$. *This is what makes the multiplicative case so different from the additive case.*

At the unregulated scale of plant $K_0$, where $R'(K_0)=i$, the expected marginal returns are $s(1-p/2)$ plus $i(1-u)(p/2)$. Unlike the additive case, this is *less than* a linear combination of $s$ and $i$. Expected marginal profits are:

$$\begin{aligned}\pi_2'(K_0) &= [s(1-p/2)+i(1-u)p/2]-i \\ &= (s-i)(1-p/2)-iup/2\end{aligned} \tag{24'}$$

The regulated monopolist will select a scale of plant larger than the unregulated scale if[6]

$$\frac{(s-i)}{i} > \frac{up/2}{1-p/2} \tag{24''}$$

A sufficiently large spread must exist between $s$ and $i$ to make it profitable to select a capital stock in excess of $K_0$, the unregulated scale.

The quasi rents must cover interest plus depreciation if the firm is to recover its investment, therefore, $i$ is considerably above the interest rate. The ratio $s/i$ must exceed the quantity $1+u(p/2)/[1-(p/2)]$ if the firm is to be induced to select a scale of plant greater than that chosen by an unregulated firm. The table below gives the greatest lower bound of $s/i$, for selected values of $u$ and $p/2$, which would satisfy inequality (24'').

| | $u = .25$ | $u = .5$ |
|---|---|---|
| $p/2 = .25$ | 1.08 | 1.17 |
| .30 | 1.11 | 1.21 |
| .40 | 1.17 | 1.33 |

If the cost of capital were 20 percent per annum, resulting from an interest rate of 10 percent and a depreciation rate of 10 percent, then the regulated rate of return must exceed 23.4 percent when $u=.25$, $p/2=.40$ or $u=.5$, $p/2=.25$. It must exceed 26.6 percent when $u=.5$, $p/2=.4$.

It is natural to call $u(p/2)/(1-p/2)$ a measure of risk. Then, the spread $(s-i)/i$ must exceed this measure of risk if regulation is to induce a larger scale of plant than would be selected by the unregulated monopolist.[7]

[6] We also require that Max $\pi_2(K)$ in equation (23) be positive.

[7] It is worthwhile to compare Perrakis' conclusions with ours. His main theme is as follows:

The firm must satisfy *ex post* (i.e., after the value of the

random demand becomes known) both the rate of return constraint and the demand constraint. . . . At some 'states of nature,' it must increase its labor above a certain minimum in order to stay within the allowed rate of return. This, however, forces it to underutilize the invested capital in order to meet demand exactly. These states of nature will occur with a positive probability whenever the firm chooses *ex ante* the output price and capital stock by maximizing profit (or expected utility of profit. . . . Barring restrictive assumptions, therefore, there is very little that can be said about relative sizes of capital stock, price, expected output, etc. [p. 415].

An explanation of his results is as follows. His firm selects both the price and the scale of plant *ex ante*. It is then required to satisfy the quantity demanded at the given price and also the rate of return constraint. Let the quantity demanded be $D(p, x)$ where $x$ is the stochastic variable. Production is $F(VK, L)$ where $V$ is the rate of utilization. The rate of return constraint is

$$sK = pD(p, x) - WL \tag{a}$$

Since the firm must also meet the demand at the price set *ex ante*,

$$F(VK, L) = D(p, x) \tag{b}$$

It follows that

$$sK = pF(VK, L) - WL \tag{c}$$

*Ex ante* controls are $p$ and $K$, and the wage is exogenous. In our analysis, only the scale of plant $K$ is an *ex ante* control; and price and labor inputs are *ex post* controls.

Demand shifts, in his analysis, are reflected by changes in $x$. When $x$ rises, total employment must rise to satisfy the rate of return constraint (a). This means that $dL/dx$ is positive. To determine the effect upon the utilization rate $V$, differentiate (c) with respect to $x$ and derive

$$(W - pF_L)\frac{dL}{dx} = pF_1K\frac{dV}{dx} \tag{d}$$

The regulated firm is a monopolist, so at the unregulated point the wage is *less than* the value of the marginal product of labor $pF_L$. The wage is equal to the marginal revenue product of labor $MR \cdot F_L < pF_L$, where $MR$ is marginal revenue. This means that $(W - pF_L)$ is negative. Consequently $dV/dx$ is negative. When $x$ rises, the employment rises and the utilization rate declines. This is how we interpret his results. Our utilization rate is constant but price is an *ex post* control. Whereas our main focus is upon the optimum *ex ante* scale of plant, he did not solve for the effect of a change in the regulated rate of return upon the optimum *ex ante* scale of plant.

## III. Social Welfare Implications of Rate of Return Regulation

Suppose that the economy were divided into two sectors: a competitive sector $X$ and a monopoly sector $Y$. Long-run equilibrium would be at point $A$ in Figure 2.

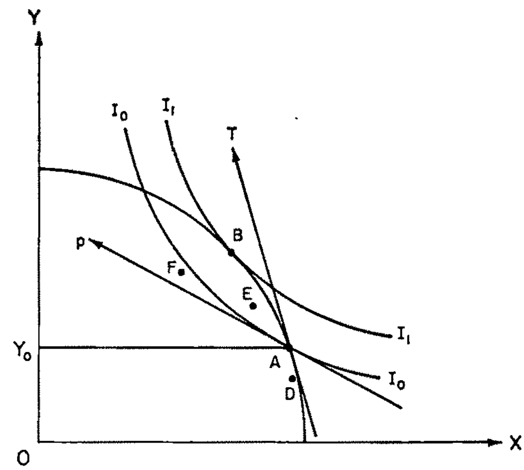



FIGURE 2. WHEN THERE IS NO REGULATION, THE ECONOMY IS AT POINT $A$. IN THE CONVENTIONAL CASES, REGULATION WILL DRIVE THE ECONOMY TO POINT $E$, AND IF IT IS TOO STRINGENT, TO POINT $F$. WHEN THE DISTURBANCE IS MULTIPLICATIVE AND THE SPREAD $(s-i)/i$ IS TOO LOW, BUT POSITIVE, REGULATION MAY DRIVE THE ECONOMY TO POINT $D$.

Within the competitive sector, price $p_X$ would be equal to both short- ($SMC_X$) and long-run ($LMC_X$) marginal cost. In the monopoly sector, the rate of output would be $Y_0$. Capital stock $K_0$ would be employed so that long-run ($LMC$) and short-run marginal cost $SMC(K_0)$ would equal marginal revenue $MR$ at that rate of output (Figure 3).

The economy would be operating on its production possibility curve because, in each sector, the rate of output is produced in the cheapest possible manner. Nevertheless, there is less than the optimal production of the monopoly output. The marginal rate of transformation in production $T = -dY/dX$ is the ratio of the long-run marginal cost of producing $X$ to that of producing $Y$. The marginal rate of substitution in consumption $p = -dY/dX$ is the ratio of the price of $X$ to the price of $Y$. In Figure 2, the marginal rate of transformation $T$ in production exceeds the marginal rate of substitution $p$ in consumption because the marginal revenue in

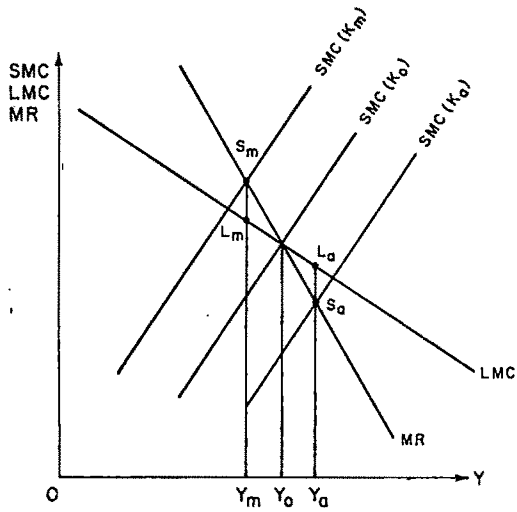



FIGURE 3. LONG-RUN MARGINAL COST IS $LMC$ AND MARGINAL REVENUE IS $MR$. THE UNREGULATED CAPITAL STOCK IS $K_0$. REGULATION MAY INDUCE THE FIRM TO SELECT CAPITAL STOCK $K_a$ OR $K_m$ DEPENDING UPON THE SPREAD $(s-i)/i$ RELATIVE TO THE RISKINESS OF THE INVESTMENT. THE CONVENTIONAL A-J EFFECT IMPLIES THAT $K_a$ IS CHOSEN. THE ANTI A-J THEOREM OPENS THE POSSIBILITY THAT $K_m$ IS CHOSEN.

sector $Y$ is less than the price: $MR_Y < P_Y$.

$$(25) \quad T = \frac{LMC_X}{LMC_Y} = \frac{P_X}{MR_Y} > \frac{P_X}{P_Y} = p$$

as described in Figure 2.

An expansion of production of sector $Y$ along the production possibility curve to point $B$ would raise social welfare to indifference curve $I_1$ from curve $I_0$.

Suppose that rate of return regulation were used to induce an expansion of the monopoly output. In the certainty case, or when inequality (21″) or (24″) is satisfied, the regulated monopolist would select a scale of plant $K_a$ in excess of $K_0$. See Figure 3. The short-run marginal cost curve associated with a capital stock $K_a$, denoted by $SMC(K_a)$ lies uniformly to the right of the short-run marginal cost curve $SMC(K_0)$ associated with the capital stock chosen by the unregulated monopolist. Consequently, the regulated monopolist would produce output $Y_a$. To be sure, there is a larger scale of output; however, production is no longer efficient. Output $Y_a$ is not being produced in the cheapest possible manner.[8] Long-run marginal cost $L_a$ exceeds short-run marginal cost $S_a$. This means that in producing output $Y_a$, the marginal capital cost $i/MPK$ (where $MPK$ is the marginal product of capital) exceeds the marginal labor cost $W/MPL$ (where $MPL$ is the marginal product of labor).

$$(26) \qquad \frac{i}{MPK} > \frac{W}{MPL}$$

Indeed the economy has been induced to produce more of the output of the monopolized sector, but it is no longer operating on its production possibility curve. In the neighborhood of the unregulated point $A$ (Figure 2), a slight decrease of $s$ below the unregulated rate $K_0M/OK_0$ in Figure 1 will drive the economy into the region between indifference curve $I_0$ and the production possibility curve[9] (say) to point $E$ (in Figure 2). The greater the degree of monopoly, the greater the gap between the marginal rate of transformation $T$ and the marginal rate of substitution $p$. Welfare will be increased by expanding the output of the monopoly sector, even though the higher rate of output is not being produced efficiently (i.e., the economy is operating inside the production possibility curve). For large deviations of $s$ from the unregulated rate of return, the economy could be driven to point $F$ where welfare is in fact reduced.

If the stochastic term were multiplicative, then the establishment of a "fair" rate of return $s$ where

$$(27) \qquad 0 < \frac{s-i}{i} < \frac{up/2}{1-p/2}$$

would violate inequality (24″). The monopolist would be induced to select *ex ante*

[8] See Stein and Borts.
[9] This is an explanation of Sheshinski's result.

a capital stock $K_m$ *below* $K_0$. The short-run marginal cost curve would be $SMC(K_m)$ in Figure 3; and the output produced would be $Y_m < Y_0$. Not only would such rate of return regulation lead to a decline in output, but the smaller output is not produced efficiently. Marginal labor cost $W/MPL$ would exceed marginal capital cost $i/MPK$. This is, of course, the reverse of the A-J effect. The net result is to drive the economy to point $D$ in Figure 2, where welfare is unambiguously lowered.

The moral of this story is that the nature of the uncertainty determines whether the A-J effect, or the anti A-J effect, occurs. Reasonable policy would aim to make the spread $(s-i)/i$ positively related to our measure of risk.

REFERENCES

H. Averch and L. L. Johnson, "Behavior of the Firm Under Regulatory Constraint," *Amer. Econ. Rev.*, Dec. 1962, *52*, 1053–69.

W. J. Baumol and A. K. Klevorick, "Input Choices and Rate-of-Return Regulation: An Overview of the Discussion," *Bell J. Econ.*, Autumn 1970, *1*, 162–90.

H. E. Leland, "Theory of the Firm Facing Uncertain Demand," *Amer. Econ. Rev.*, June 1972, *62*, 278–91.

S. Perrakis, "On the Regulated Price-Setting Monopoly Firm with a Random Demand Curve," *Amer. Econ. Rev.*, June 1976, *66*, 410–16.

A. Sandmo, "On the Theory of the Competitive Firm under Price Uncertainty," *Amer. Econ. Rev.*, Mar. 1971, *61*, 65–73.

E. Sheshinski, "Welfare Aspects of a Regulatory Constraint," *Amer. Econ. Rev.*, Mar. 1971, *61*, 175–78.

J. L. Stein and G. H. Borts, "Behavior of the Firm Under Regulatory Constraint," *Amer. Econ. Rev.*, Dec. 1972, *62*, 964–70.

# The Benefits and Costs of Rate of Return Regulation

By JEFFREY CALLEN, G. FRANK MATHEWSON, AND HERBERT MOHRING*

The seminal article "Behavior of the Firm Under Regulation" by Harvey Averch and Leland Johnson has evoked extensive comment during the years since it appeared. Only in recent articles by Eytan Sheshinski and Alvin Klevorick, however, has the Averch-Johnson analytical framework been employed to deal with the question: If, for whatever reason, a rate of return constraint is the device that must be employed to control a monopolist's profits, at what level should the allowable rate of return be set? The answer to this question and pseudoempirical analysis of its implications are the subjects of this paper.

Specifically, Section I provides a simpler development of Sheshinski's and Klevorick's main formal conclusions than was contained in their papers.[1] Section II analyzes, as did Klevorick, the special case of constant elasticity demand and Cobb-Douglas production functions in a world in which income distribution considerations can be ignored and in which the regulatory process involves no resource costs. Expressions are derived and tabulated for the optimum allowable rate of return and for the effects on "social welfare" (i.e., the sum of consumers' and producers' surpluses), output, costs, and capital intensity of imposing a rate of return constraint on a formerly unregulated monopoly.

If this simple example sustains generalization, an allowable rate of return in excess of a monopoly's cost of capital would be desirable only if some combination prevails of a low demand elasticity, modest scale economies, and a small exponent on capital in the production function.[2] Furthermore, the gains from rate of return regulation (gross of the costs of the regulatory process) are substantial even if the allowed return differs considerably from its optimum value. Finally, at least for combinations of parameter values which reflect capital intensive production processes with substantial scale economies, rate of return regulation could capture a substantial share of the benefits derivable from setting price equal to average or (except when the elasticity of demand is large) even marginal cost.

## I. The Optimum Allowable Rate of Return

As preparation for dealing with the optimization problem involved in specifying the rate of return allowed a regulated monopoly, consider first the optimization problem faced by the monopoly itself. Its profits can be written:

$$(1) \qquad \Pi = R(Q) - wL(Q, K) - rK$$

where $R(Q)$ is the revenue derived from selling $Q$ units of product; $L(Q, K)$—obtained by inverting the production function—is the amount of variable input re-

* Assistant professor of business economics, University of Manitoba; associate professor of economics, University of Toronto; and professor of economics, University of Minnesota, respectively. Most of the work described in this paper was undertaken during 1973 while we were all associated with the University of Toronto's Faculty of Management Studies and department of political economy. We are indebted to Bela Balassa and Carl Christ for helpful suggestions.

[1] Albeit, in Sheshinski's case, at the expense of employing a less general social welfare function.

[2] Ignoring the motivational and indeterminacy problems that arise when the allowable rate of return is set equal to the cost of capital.

quired to produce $Q$ units of output if a capital stock valued at $K$ dollars is employed; $w$ is the wage rate of the variable input; and $r$ is the cost of capital. Depreciation expenses are assumed away.

The constraint that the monopoly is allowed to earn a rate of return $s$, which is less than the rate $s^*$ it would earn in the absence of regulation, can be written

$$(2) \qquad sK = R(Q) - wL(Q, K)$$

This constraint can be interpreted as specifying the required capital stock $K$ as an implicit function of the output level $Q$. Using this implicit function to eliminate $K$ in equation (1) reduces the set of decision variables to $Q$ alone. Proceeding in this fashion avoids the difficulties involved in interpreting auxiliary variables. Differentiating the amended equation (1) with respect to $Q$ yields

$$(3) \quad R' - w(L_Q + L_K K') - rK' = 0$$

where, from (2)

$$sK' = R' - w(L_Q + L_K K')$$

or:[3]

$$(4) \qquad K' = (R' - wL_Q)/(s + wL_K)$$

Substituting (4) into (3) and rearranging terms yields:

$$(5) \qquad (s - r)(R' - wL_Q) \doteq 0.$$

If $r$ is less than $s$, this reduces to:

$$(6) \qquad R'(Q) = wL_Q(Q, K)$$

As several writers have noted, if $s = r$, the behavior of the monopoly is indeterminate since each of the continuum of $K$ values which satisfies equation (2) would yield the same (zero) economic profit.

Equation (6) is the analogue for a regulated monopoly of the standard condition for profit maximization, marginal revenue equals marginal cost: $R'$ is marginal revenue while $wL_Q$ is regulated or pseudomarginal cost. That is, $wL_Q$ is the cost of the variable inputs required to produce an additional unit of output when $K$ is set at the level required to satisfy the rate of return constraint, equation (2). In the range $r < s < s^*$, using equation (2) to eliminate $K$ from equation (6) yields $Q^{**} = Q^{**}(s, w)$, the monopoly's profit-maximizing output as a function of the allowable rate of return and the wage rate of the variable input. Substituting $Q^{**}$ for $Q$ in equation (2) yields the profit-maximizing capital input level $K^{**} = K^{**}(s, w)$.

Turning to the regulatory authority's optimization problem, we assume, as did Klevorick, the social objective to be maximization of the sum of consumers' and producers' surpluses:

$$(7) \quad W = \int_0^{Q^{**}} P(q)dq - wL(Q^{**}, K^{**}) - rK^{**}$$

Differentiating with respect to $s$ yields:

$$(8) \qquad [P - wL_Q]Q_s^{**} = [r + wL_K]K_s^{**}$$

as the necessary condition for maximizing $W$. The left-hand side of this equation is the difference between price and (pseudo) marginal cost times the change in output resulting from a change in the allowable rate of return—the marginal net benefit of a change in $s$. The right-hand side of equation (8) is the change in capital employed by the monopoly as a result of a change in $s$ times the difference between the annual cost of a dollar's worth of capital plant and the reduction in variable input costs resulting from employment of that additional unit of capital. That is, the right-hand side of (8) equals the marginal

[3] Cost minimization would dictate employing that amount of capital for which $r + wL_K = 0$. If the monopolist is to stay in business, $s$ must be greater than or equal to $r$. As is well known from the Averch-Johnson literature, a rate of return constraint induces the employment of an inefficiently large amount of capital, i.e., a level for which $r + wL_K > 0$. The denominator on the right-hand side of (4) would therefore be positive.

net increase in costs resulting from the excessive substitution of capital for labor that rate of return regulation induces.

An unconstrained profit-maximizing monopoly would produce at minimum cost. Therefore, with $s=s^*$, $wL_Q$ would equal *the* marginal cost of its product while $r$ would equal $-wL_K$. This being the case, when evaluated at $s=s^*$, the left-hand side of (8) would be positive while the right-hand side would be zero. Reducing $s$ to a value less than $s^*$ could be expected to lower the difference between price and (pseudo) marginal cost. Also, by inducing the substitution of capital for labor, reducing $s$ would lead $-wL_K$ to fall short of $r$. These considerations suggest that if administration costs could be ignored, reducing $s$ below $s^*$ would increase our measure of social welfare. That is, they suggest that regardless of the specific natures of the production and demand functions involved, the optimum allowable rate of return is less than that which an unconstrained monopolist would earn. It also *seemed* plausible to conjecture that the optimum value of $s$ could be shown to lie in the range $r<s<s^*$. We were unable to demonstrate the general validity of this conjecture, however.[4] Indeed, the specific example described in the following section provides a counterexample to it, at least when increasing returns to scale are involved.

## II. The Benefits and Costs of Rate of Return Regulation: An Example

We suppose, as did Klevorick, that the monopoly is faced by a constant elasticity demand schedule and that its production function is Cobb-Douglas. Specifically, the demand function is:

$$P = AQ^{-\epsilon} \tag{9}$$

where $A$ is a constant and $1/\epsilon$ is the elasticity of demand. By appropriately specifying the units in which output is measured, the production function can be written without further loss of generality as

$$Q^{\beta} = K^{\alpha}L^{1-\alpha} \tag{10}$$

where $0<\alpha<1$, and $\beta\leq 1$ is the reciprocal of the order of homogeneity of $Q$. For notational convenience, it proves useful to define the following combinations of parameters:

$$\begin{aligned}
\gamma &\equiv \beta + \epsilon - 1 \\
\delta &\equiv w\beta/[A(1-\epsilon)(1-\alpha)] \\
\eta &\equiv \gamma + \alpha(1-\epsilon) \\
&\equiv \beta - (1-\epsilon)(1-\alpha)
\end{aligned}$$

In the absence of constraints, the monopoly's profit function would be

$$\Pi = AQ^{1-\epsilon} - wL - rK \tag{11}$$

Using (10) to eliminate $Q$ in equation (11) and differentiating with respect to $L$ and $K$ yields[5] the profit-maximizing values of these variables. Substituting the results into equation (10) yields the profit-maximizing output level:

$$Q^* = [A(1-\epsilon)(\alpha/r)^{\alpha}((1-\alpha)/w)^{1-\alpha}/\beta]^{1/\gamma} \tag{12}$$

The difference between the revenue derived from this output level and the cost $wL^*$ of the variable input used in producing it is the monopoly's accounting profit. This difference divided by its profit-maximizing capital stock yields the ratio of its rate of return on invested capital as conventionally measured to its cost of capital:

$$s^*/r = [\beta/(1-\epsilon) - (1-\alpha)]/\alpha \tag{13}$$

The profit-maximizing capital stock itself can be written:

$$K^* = [\alpha w/((1-\alpha)r)]^{1-\alpha}Q^{*\beta} \tag{14}$$

[4] Roger Sherman (pp. 390–93) presents a geometric analysis suggesting the validity of this conjecture when the monopoly operates under constant returns to scale.

[5] The algebra involved is straightforward if tedious. Only the final results and a sketchy outline of the steps involved will therefore be given.

Substituting $K^*$ and $L^*$ into $C=wL+rK$ yields the monopoly's total cost:

$$(15)\quad C^* = (r/\alpha)^{\alpha}(w/(1-\alpha))^{1-\alpha}Q^{*\beta}$$

Turning to the regulated monopoly, substituting the demand and production functions into equation (2) and using the result to eliminate $K$ in equation (6) yields its profit-maximizing output:

$$(16)\quad Q^{**} = [(A\delta - w)/(s\delta^{1/\alpha})]^{\alpha/\gamma}$$

Inserting (16) into equation (2) yields the profit-maximizing capital stock:

$$(17)\quad K^{**} = \delta^{(1-\alpha)/\alpha}Q^{**\eta/\alpha}$$

Introducing these last two equations into $L=L(Q, K)$ and the result into $C=wL+rK$ yields the constrained monopoly's total cost function:

$$(18)\quad C^{**} = wQ^{**1-\epsilon}/\delta + r\delta^{(1-\alpha)/\alpha}Q^{**\eta/\alpha}$$

For the Cobb-Douglas production function under examination, $L(Q, K)$ is:

$$(19)\quad L = (Q^{\beta}/K^{\alpha})^{1/(1-\alpha)}$$

Inserting equation (9), the derivatives of $L$ with respect to $K$ and $Q$, and the derivatives of equations (16) and (17) with respect to $s$ into equation (8) yields the optimum allowable rate of return $s^{opt}$ *if* a value of $s>r$ satisfies equation (8):

$$(20)\quad s^{opt} = \eta^2 r/[\alpha(\beta - (1-\alpha)(1-\epsilon)^2)]$$

If this equation is satisfied only for a value of $s \leq r$, welfare maximization would involve a corner solution. That is, it would require establishing an $s$ sufficiently in excess of $r$ to avoid the indeterminacy (see equation (5)) which arises when $s$ equals $r$ and the incentive to withdraw from production associated with an $s<r$. Note that the ratio $s^{opt}/r$ derivable from equation (20) depends only on parameters having straightforward economic interpretations: the elasticity of demand $(1/\epsilon)$, the exponent of capital in the production function $(\alpha)$, and the order of homogeneity of that function $(1/\beta)$.

Division of equation (16) by equation (12) gives the relative increase in output effected by establishing a ratio $r/s$ of the cost of capital to the allowable rate of return as a function of these same parameter values:

$$(21)\quad Q^{**}/Q^* = [(r/s)\eta/\alpha(1-\epsilon)]^{\alpha/\gamma}$$

Substitution of $Q^{**}$ for $Q^*$ in equation (14) yields the amount of capital that would be required for production of $Q^{**}$ at minimum cost. Division of the resulting expression into equation (17) then yields a measure of the extent to which rate of return regulation leads to excessive employment of capital:

$$(22)\quad K^{**}(Q^{**})/K^*(Q^{**}) = [(r/s)\eta/(\alpha(1-\epsilon))]^{1-\alpha}$$

A similar substitution in equation (15) and division into equation (18) yields a measure of the cost increasing effect of rate of return regulation:

$$(23)\quad C^{**}(Q^{**})/C^*(Q^{**}) = (1-\alpha)[(s/r)\alpha(1-\epsilon)/\eta]^{\alpha} + \alpha[(r/s)\eta/(\alpha(1-\epsilon))]^{1-\alpha}$$

Given the way the analysis has been set up, the logical measure of the social benefits attributable to rate of return regulation might seem to be $W^{**}/W^*$, "net social welfare" contributed by the monopoly's operations under regulation divided by that associated with its unregulated behavior. However, as the absolute value of the elasticity of demand approaches one from above, the area under a constant elasticity demand schedule approaches infinity. This being the case, the *percentage* gain in $W$ resulting from a given *absolute* gain in the sum of consumers' and producers' surpluses diminishes with reductions in the elasticity of demand. With low elasticity values, small percentage

Table 1—Rate of Return, Output, Cost, and Welfare Gain Ratios for Alternative Capital Exponents, Scale Economy Coefficients, and Demand Elasticities

| $\alpha$ | $\beta$ | $1/\epsilon$ | $s^*/r$ | $s^{opt}/r$ | $Q^{**}/Q^*$ | $K^{**}/K^*$ | $C^{**}/C^*$ | $\Delta W^r/R^*$ | $\Delta W^r/\Delta W^a$ | $\Delta W^r/\Delta W^m$ |
|---|---|---|---|---|---|---|---|---|---|---|
| .33 | .6 | 1.1 | 17.8 | 1.47 | 5.12 | 5.28 | 2.05 | 1.09 | .34 | .31 |
| | | 2.0 | 1.6 | .49 | 4.79 | 1.37 | 1.03 | 1.02 | .77 | .11 |
| | .8 | 1.1 | 24.4 | 2.06 | 3.19 | 5.19 | 2.02 | .76 | .32 | .32 |
| | | 2.0 | 2.8 | 1.03 | 3.03 | 1.95 | 1.13 | .40 | .49 | .40 |
| | | 4.0 | 1.2 | .64 | 3.37 | 1.13 | 1.00 | .43 | .90 | .07 |
| | 1.0 | 1.1 | 31.0 | 2.66 | 2.46 | 5.14 | 2.01 | .58 | .32 | .32 |
| | | 2.0 | 4.0 | 1.60 | 1.84 | 1.84 | 1.11 | .20 | .39 | .39 |
| | | 4.0 | 2.0 | 1.20 | 1.98 | 1.41 | 1.03 | .11 | .54 | .54 |
| .67 | .6 | 1.1 | 9.4 | .82 | 18.81 | 2.11 | 1.48 | 2.21 | .70 | .64 |
| | | 2.0 | 1.3 | .55 | 5.75 | 1.09 | 1.01 | 1.23 | .94 | .13 |
| | .8 | 1.1 | 12.7 | 1.11 | 9.85 | 2.25 | 1.57 | 1.55 | .66 | .65 |
| | | 2.0 | 1.9 | .84 | 4.16 | 1.24 | 1.04 | .67 | .82 | .66 |
| | | 4.0 | 1.1 | .75 | 3.56 | 1.03 | 1.00 | .47 | .97 | .08 |
| | 1.0 | 1.1 | 16.0 | 1.41 | 5.92 | 2.24 | 1.56 | 1.18 | .65 | .65 |
| | | 2.0 | 2.5 | 1.14 | 2.86 | 1.30 | 1.06 | .36 | .72 | .72 |
| | | 4.0 | 1.5 | 1.04 | 2.67 | 1.13 | 1.01 | .17 | .82 | .82 |

*Note:* $\alpha\equiv$capital exponent; $\beta\equiv$scale economy coefficient; $1/\epsilon\equiv$demand elasticity; $s^*/r\equiv$monopoly rate of return $\div$ cost of capital; $s^{opt}/r\equiv$optimum allowable rate of return $\div$ cost of capital; $Q^{**}/Q^*\equiv$output ratio; $K^{**}/K^*\equiv$capital ratio; $C^{**}/C^*\equiv$cost ratio; $\Delta W^r/R^*\equiv$regulated welfare gain; $\Delta W^r/\Delta W^a\equiv$regulated welfare gain $\div$ average cost pricing welfare gain; $\Delta W^r/\Delta W^m\equiv$regulated welfare gain $\div$ marginal cost pricing welfare gain.

gains would result even from very large absolute gains. This being the case, the $W^{**}/W^*$ measure is less than ideal. An alternative measure is the change in net social welfare resulting from regulation $\Delta W^r \equiv W^{**} - W^*$, divided by the market value of the monopolist's output under unregulated conditions.

With a constant elasticity demand schedule, the net social welfare associated with $Q$ units of the monopolist's output can be written:

$$W = AQ^{1-\epsilon}/(1-\epsilon) - C(Q) \tag{24}$$

Substituting equations (12) and (15) into equation (24) yields $W^*$ while substitution of equations (16) and (18) yields $W^{**}$. Subtracting and dividing by $R(Q^*) = A(Q^*)^{1-\epsilon}$ results in:

$$\begin{aligned}\Delta W^r/R^* = {} & [1/(1-\epsilon) - (1-\epsilon)(1-\alpha)/\beta] \\ & \cdot (Q^{**}/Q^*)^{1-\epsilon} - 1/(1-\epsilon) \\ & + (1-\epsilon)/\beta - (r/s)^{\eta/\gamma}(\eta/\beta) \\ & \cdot [\eta/\alpha(1-\epsilon)]^{\alpha(1-\epsilon)/\gamma}\end{aligned} \tag{25}$$

Table 1 gives the ratios of $s^*$, the unconstrained monopoly rate of return, and $s^{opt}$, the optimum allowable rate of return (more accurately, the value of $s$ which satisfies equation (20)) to $r$ the cost of capital for alternative values of $\alpha$, $\beta$, and $\epsilon$ that yield stable equilibria.[6] In addition, it provides values of equations (21)–(23) and (25) with $s/r$ equal to $s^{opt}/r$ when $s^{opt} > r$ and $s/r$ equal to one when a value of $s \leq r$ satisfies equation (20). The equation values associated with $s/r = 1$ would not, of course, be observed in the real world. Still, it seems better to tabulate them than values associated with one plus some arbitrarily chosen increment.

[6] If $\beta + \epsilon < 1$, the demand schedule intersects the marginal cost schedule from below. Entries in this table therefore involve only $\beta + \epsilon$ values greater than one.

Perhaps the most important conclusion to be drawn from Table 1 is that the optimum allowable rate of return exceeds the cost of capital only when some combination of modest scale economies, an elasticity of demand close to one, and a small exponent of capital in the production function are in effect. If the $\Delta W^r/R^*$ benefit measure is accepted as reasonable, the potential gains (again, gross of administrative costs) from rate of return regulation are substantial. This form of regulation does produce substantial increases in capital inputs and, to a lesser extent, production costs. However, these cost increases are far more than offset by the additional consumers' plus producers' surpluses generated by the increased output. While the smallest value of the $\Delta W^r/R^*$ index tabulated (for $\alpha=0.33$, constant returns to scale, and a demand elasticity of four) is 11 percent, one index value in excess of 200 percent appears while values in excess of 100 percent are common.

Showing that rate of return regulation could make substantial contributions to social welfare given appropriate parameter combinations is not, of course, the end of the story. If raising subsidies involves no resource misallocation, the sum of consumers' and producers' surpluses would be maximized by establishing that output level for which price equals long-run marginal cost. If, for whatever reason, subsidies cannot be paid, the output level at which price equals long-run average cost is the best that can be hoped for. If the benefits derivable from rate of return regulation are close to those attainable with average or marginal cost pricing, the search for alternative regulatory procedures would be pointless.

To explore the potential benefits of alternative regulatory procedures, define the increases in net social welfare that would result from replacing unconstrained monopoly pricing by price equal to average and marginal cost, respectively, as $\Delta W^a = W^a - W^*$ and $\Delta W^m = W^m - W^*$ where $W^a$ and $W^m$ are the respective values of net social welfare resulting from setting price equal to marginal and to average cost. Using equations (9) and (15) to determine the output levels that would result from average and marginal cost pricing, making appropriate substitutions in equation (24), and dividing by $R^*$ yields:

$$(26)\quad \Delta W^m/R^* = (1-\epsilon)^{-\beta/\gamma} - (1-\epsilon)^{-(1-\epsilon)/\gamma} - \gamma$$

$$(27)\quad \Delta W^a/R^* = \beta^{(1-\epsilon)/\gamma}[\Delta W^m/R^* + \gamma] - \gamma$$

Dividing equations (26) and (27) into equation (25) yields $\Delta W^r/W^a$ and $\Delta W^r/\Delta W^m$—the ratios of benefits derivable from rate of return regulation to those associated with average and marginal cost pricing. These ratios are given in the final two columns of Table 1.

The values of these ratios differ considerably from one set of parameter values to another. The $\Delta W^r/\Delta W^a$ and $\Delta W^r/\Delta W^m$ measures range from respective lows of 0.32 and 0.07 to respective highs of 0.94 and 0.82. Both of these ratios increase with increases in the exponent of capital in the production function. Also, $\Delta W^r/\Delta W^a$ increases with increases in the elasticity of demand and in scale economies whereas $\Delta W^r/\Delta W^m$ exhibits an irregular pattern of change with increases in these variables.

It seems highly unlikely that any regulatory authority anywhere has the information at its disposal that would be necessary for precise determination of optimum allowable rates of return for the activities it controls. Indeed, with such complete data, rate of return regulation would be pointless. No less information is necessary to determine $s^{opt}$ than to dictate the inputs necessary to produce efficiently that output for which demand price equals average or marginal cost. That is, the information necessary to determine $s^{opt}$ would permit

TABLE 2— $\Delta W/R^*$ WELFARE GAIN RATIOS FOR ALTERNATIVE NONOPTIMAL RATIOS OF THE ALLOWABLE RATE OF RETURN TO THE COST OF CAPITAL

| | | | | $\Delta W^r/R^*$ when $s/r$ equals: | | | | | $\Delta W^r/R^*$ when $a$ equals: | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ | $\beta$ | $1/\epsilon$ | $s^{opt}/r$ | 1.1 | 1.25 | 1.5 | 2.0 | $a^{opt}$ | 0 | .25 | .5 | .75 |
| .33 | .6 | 1.1 | 2.06 | .58 | .63 | .68 | .73 | .03 | 1.02 | .69 | .37 | .16 |
| | | 2.0 | 1.03 | .78 | .48 | .11 | [a] | −.85[b] | 1.02 | .67 | .39 | .17 |
| | .8 | 1.1 | 2.06 | .63 | .68 | .73 | .76 | .05 | .58 | .50 | .27 | .75 |
| | | 2.0 | 2.80 | .39 | .37 | .31 | .18 | .02 | .40 | .33 | .21 | .10 |
| | | 4.0 | 1.20 | .17 | [a] | [a] | [a] | −1.82[b] | .43 | .29 | .17 | .08 |
| | 1.0 | 1.1 | 2.66 | .36 | .43 | .50 | .56 | .06 | .30 | .40 | .21 | .09 |
| | | 2.0 | 1.60 | .13 | .17 | .19 | .18 | .20 | .09 | .19 | .14 | .07 |
| | | 4.0 | 1.20 | .11 | .11 | .08 | [a] | .20 | .08 | .11 | .08 | .04 |
| .67 | .6 | 1.1 | .82 | 2.17 | 2.11 | 1.99 | 1.75 | −.02[b] | 2.21 | 1.30 | .71 | .30 |
| | | 2.0 | .55 | .69 | .14 | [a] | [a] | −1.52[b] | 1.23 | .81 | .48 | .21 |
| | .8 | 1.1 | 1.11 | 1.55 | 1.54 | 1.50 | 1.39 | .01 | 1.54 | .96 | .52 | .22 |
| | | 2.0 | .84 | .59 | .47 | .27 | [a] | −.18[b] | .67 | .49 | .31 | .14 |
| | | 4.0 | .74 | [a] | [a] | [a] | [a] | −2.59[b] | .47 | .31 | .19 | .08 |
| | 1.0 | 1.1 | 1.41 | 1.15 | 1.17 | 1.18 | 1.13 | .03 | 1.12 | .76 | .41 | .18 |
| | | 2.0 | 1.14 | .36 | .35 | .30 | .14 | .09 | .34 | .33 | .22 | .11 |
| | | 4.0 | 1.04 | .16 | .11 | [a] | [a] | .08 | .17 | .16 | .11 | .06 |

*Note:* $\alpha\equiv$ capital exponent; $\beta\equiv$ scale economy coefficient; $1/\epsilon\equiv$ demand elasticity; $s^{opt}/r\equiv$ optimum allowable rate of return ÷ cost of capital; $s/r\equiv$ allowable rate of return ÷ cost of capital; $a\equiv(s-r)/(s^*-r)\equiv$ deviation of allowable rate of return from cost of capital ÷ deviation of monopoly rate of return from cost of capital.

[a] $s/r>s^*/r$.

[b] Equation (20) satisfied for $s<r$.

complete elimination of Averch and Johnson-type distortions.

For this reason, it is of interest to determine how the benefits suggested by Table 1 would be affected by specification of a nonoptimum allowable rate of return. It is this issue to which Table 2 is addressed. It lists $\Delta W^r/R^*$ for values of $s/r$ different from the optimum. In the first group of columns, ratios are calculated for allowable rates of return equal, alternatively, to 1.1, 1.25, 1.5, and 2 times the cost of capital when these values do not exceed the rate of return an unconstrained monopolist would earn. In the second group of columns, ratios are calculated for $s/r$ values equivalent, alternatively, to $a\equiv(s-r)/(s^*-r)$ values of 0, 0.25, 0.5, and 0.75. The ratio $a$ is the amount by which the allowable rate of return exceeds the cost of capital divided by the amount by which the unconstrained monopoly rate of return exceeds the cost of capital.

Deviations from optimum allowable rates of return do lead to reduced benefits, of course. Reductions are particularly great when allowable rates are close to the cost of capital under circumstances in which $s^{opt}/r$ is substantially greater than one. But the benefits listed in Table 2 are positive and typically substantial even when there is a considerable difference between the allowable and the optimum rate of return. If—a very big if—generalization is warranted on the basis of these tables, regardless of the values of the technological and demand parameters under which it operates, subjecting a

formerly unregulated monopoly to the constraint that its rate of return on invested capital not exceed, say, 1.1 times its cost of capital would generate social gains substantially in excess of those which would accrue in the absence of regulation. Indeed, for some combinations of values of the scale economy, capital intensity, and demand elasticity variables, this form of regulation could yield benefits almost as great as those derivable from cost minimizing production coupled with average or even marginal cost pricing.

The literature contains a number of reports on attempts to fit Cobb-Douglas production functions to regulated utilities. Unfortunately, however, we have found only one such study which provides estimates of all three of the parameters that enter into our model. In this study, A. Rodney Dobell et al. derived estimates of our $\alpha$ and $\beta$ parameters equal to 0.36 and 0.90 respectively from an analysis of Bell Canada data for the period 1952–67. In addition, when evaluated for 1967 conditions, their demand relationship yields a long-run price elasticity of 2.01.

From equation (13), an unconstrained profit-maximizing monopolist subject to these conditions would earn a rate of return equal to 3.2 times its cost of capital. From equation (20), the optimum ratio of the allowable rate of return to the cost of capital would be 1.26. Imposing this constraint would generate a $\Delta W^r/R^*$ benefit (equation (25)) of 0.28. This benefit is 45 and 43 percent respectively of those that would result from average and marginal cost pricing (equations (26) and (27)). Setting the allowable rate of return equal to 1.1 times the cost of capital would only lower the $\Delta W^r/R^*$ benefit to 0.27, 43, and 41 percent respectively of the average and marginal cost pricing benefits. Even allowing a rate of return equal to twice the cost of capital would generate benefits equal to 30 and 29 percent respectively of those associated with average and marginal cost pricing.

## REFERENCES


H. Averch and L. L. Johnson, "Behavior of the Firm Under Regulation," *Amer. Econ. Rev.*, Dec. 1962, *52*, 1053–69.

A. R. Dobell et al., "Telephone Communications in Canada: Demand, Production, and Investment Decisions," *Bell J. Econ.*, Spring 1972, *3*, 175–219.

A. K. Klevorick, "The 'Optimal' Fair Rate of Return," *Bell J. Econ.*, Spring 1971, *2*, 122–53.

R. Sherman, *The Economics of Industry*, Boston 1974.

E. Sheshinski, "Welfare Aspects of a Regulatory Constraint: Note," *Amer. Econ. Rev.*, Mar. 1971, *61*, 175–78.

# The Economics of Multiple Job Holding

*By* Robert Shishko and Bernard Rostker*

A person holding two or more jobs is said to be moonlighting, or participating in a secondary labor market.[1] This study investigates the determinants of the moonlighting supply function in terms of demographic and market factors and describes the relationship between primary and secondary employment.

Economic literature has treated moonlighting in two ways. First, there have been several attempts to extend traditional micro-economic theory to explain the individual moonlighter's supply curve.[2] Second, some researchers have presented demographic profiles of the typical moonlighter.[3] To our knowledge, no one has combined these two approaches to estimate a moonlighting supply curve. In this paper we attempt to estimate the moonlighting supply curve with data from the Income Dynamics Panel (*IDP*) using the Tobit technique for estimating relationships with limited dependent variables.

## I. The Economics of Moonlighting

Traditionally, an individual receives purchasing power, or income, as payment for work. Time spent to obtain this income can be viewed as forgone leisure. In Figure 1, which shows a representative set of indifference curves indicating equal-utility combinations of income and leisure, point $B$ represents the maximum amount of leisure available per period. The slope of the line $AB$ is the negative of the prevailing wage rate, i.e., the rate at which leisure can be traded for income in the labor market. If an individual is assumed to maximize utility, he will choose a contract such that the prevailing wage rate is equal to his marginal rate of substitution of leisure for income. Such a contract is point $P$, the point of tangency between the indifference curve $\Omega_2$ and the wage line $AB$. The locus of all such tangency points $SS'$, i.e., the price expansion path, can easily be translated into the individual supply curve.

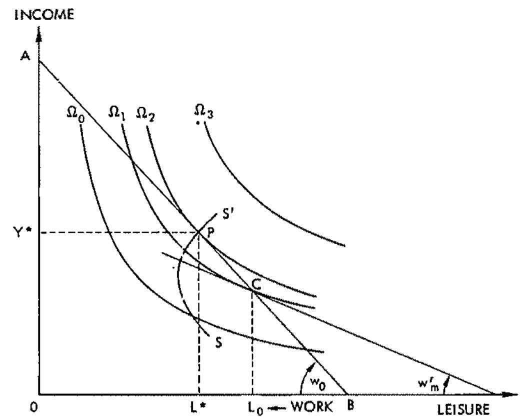



Figure 1. Utility Maximization in Income and Leisure with and without Restricting Hours Worked on Primary Job

### A. *Multiple Job Holding—A Geometric Treatment*

An individual's willingness to take a second job depends on whether he can

* The Rand Corporation, Santa Monica. We wish to acknowledge the invaluable assistance provided by Rand colleagues John Cogan, Glenn Gotz, Charles Phelps, C. Robert Roll, Finis Welch, and our dedicated programmer, Laura Sammis. Views expressed in this paper are our own, and are not necessarily shared by Rand or its research sponsors.

[1] The term secondary employment has also been used in the literature to refer to employment by groups who enter and withdraw from the labor market at irregular intervals, e.g., married women. In this paper we refer to secondary employment as the holding of a job or jobs in addition to steady, full-time employment.

[2] For example, see Richard Perlman and Leon Moses.

[3] For example, see Harold Guthrie (1965, 1969), Harvey Hamel, and Allyson Grossman.

work enough hours at his prevailing primary wage rate to satisfy his income goals. Consider an individual whose primary job allows him to work only $L_0$ hours but who would like to work $L^* - L_0$ additional hours. This restriction forces the individual to indifference curve $\Omega_1$ at point $C$. Ignoring any additional costs of securing the second job, or any additional job-related costs (such as transportation), an individual will accept a second job as long as it pays a wage above the marginal rate of substitution of income for leisure at point $C$—the intersection of the primary wage line and the allowable hours on the first job.[4] If the wage rate on the second job lies between this minimum, $w_m^r$, and the primary wage rate $w_0$, he will take a second job, and the total hours worked will be somewhat fewer than the number he desired to work on his first job. If the wage rate on the second job exceeds that on his first job, depending on his preference pattern, he may decide to work even more hours than he originally desired. Moreover, if there are no limits on the amount of time he can spend on the second job, he may even substitute it for his primary employment—as sometimes occurs when people make their avocation their vocation. However, the general character of second jobs often limits the number of hours that can be worked to less than "full time" (for example, seasonal work), or is contingent upon or complementary to the primary employment (for example, when a school teacher tutors students after class), or has an unacceptable uncertainty of income given the person's risk aversion. If an individual is completely free to determine the number of hours he wants to work, at a high enough secondary wage he may develop a backward bending supply curve; and an increase in his secondary wage might result in a decrease in the number of hours worked.

Changes in the primary wage alter the minimum wage necessary to induce people to take a second job. In theory, an increase in the primary wage rate can result in an increase or a decrease in the minimum acceptable second-job wage rate. Moreover, an increase in the primary wage can increase or decrease the hours offered on the secondary labor market.

Just as a change in the primary wage rate affects both the reservation wage for moonlighting and the hours offered in the secondary labor market, so will a change in the primary job hours affect these two variables. An increase in the first-job hours $L_0$ can result in an increase or a decrease in the moonlighting reservation wage, though a large increase in $L_0$ will probably result in an increase in that wage. An increase in $L_0$ can result in an increase or a decrease in the moonlighting hours offered to the market, depending only on whether the moonlighting wage is greater or less than the primary job wage, and on whether leisure is a superior good. If the primary wage rate exceeds the moonlighting wage rate *and* leisure is a superior good, moonlighting hours offered will unambiguously decrease. If the primary wage rate equals the moonlighting wage rate, then the situation depicted in Figure 1 prevails, and moonlighting hours offered will be decreased on a one-for-one basis as primary hours increase. Only if the primary wage rate is less than the moonlighting wage rate could an increase in primary hours result in an increase in moonlighting hours offered.

### B. *Multiple Job Holding—A Mathematical Treatment*

Although geometric techniques are useful in explaining economic relationships,

[4] The minimum second-job wage rate above which the individual will accept secondary employment is called the second-job reservation wage. Let $w_m^r(w_0, L_0)$ designate this reservation wage when $w_0$ and $L_0$ are the first-job wage and first-job hours, respectively.

a more precise formulation can be presented mathematically. Consider a representative individual with a utility function embodying a general consumption good $x$, and leisure $l$. Suppose this person holds a primary job that contractually calls for $L_0$ hours of work at a fixed wage rate $w_0$. Assume further that he is underemployed in this situation (as in Figure 1) and seeks $L_m$ hours of secondary employment. The magnitude of $L_m$ depends upon the moonlighting wage rate $w_m$, which the job holder is powerless to affect. The individual then maximizes $U(x, l)$ subject to

$$(1) \qquad A_0 + w_0L_0 + w_mL_m - x \geq 0$$

$$(2) \qquad N - L_m - L_0 - l \geq 0$$

$$(3) \qquad x, l, L_m \geq 0$$

where the price of $x$ is taken as the numeraire, $N$ is the fixed number of hours per unit of time available for work or leisure, and $A_0$ is nonlabor income.

Define the function[5]

$$\psi(w_m; Y_0) = \max U(x, l)$$

subject to

$$(4) \qquad Y_0 - x - w_ml \geq 0$$
$$x, l \geq 0$$

where

$$(5) \qquad Y_0 = w_0L_0 + w_m(N - L_0) + A_0$$

Then the supply function for moonlighting labor is

$$(6) \qquad L_m = N - L_0 + \frac{\partial \psi}{\partial w_m} \bigg/ \frac{\partial \psi}{\partial Y_0}$$

The third term of the right-hand side of equation (6) is the negative quantity of leisure demanded at price $w_m$. Substituting this expression for $l$ in equation (2), which holds with equality at an optimum, the supply of moonlighting labor is obtained as the complement of the demand for leisure. The properties of this supply function can be uncovered by totally differentiating the first-order conditions,[6] equations (7)–(9), to obtain (10).

$$(7) \qquad U_l - U_xw_m = 0$$

$$(8) \qquad A_0 + w_0L_0 + w_mL_m - x = 0$$

$$(9) \qquad N - L_0 - L_m - l = 0$$

$$(10) \qquad \Lambda \begin{bmatrix} dL_m \\ dl \\ dx \end{bmatrix} = J$$

where

$$\Lambda = \begin{bmatrix} w_m & 0 & -1 \\ -1 & -1 & 0 \\ 0 & (w_mU_{xl} - U_{ll}) & (w_mU_{xx} - U_{xl}) \end{bmatrix}$$

and

$$J = \begin{bmatrix} -dA_0 - w_0dL_0 - L_0dw_0 - L_mdw_m \\ dL_0 \\ -U_xdw_m \end{bmatrix}$$

The second-order conditions at equilibrium guarantee that $\Lambda > 0$, since

$$\Lambda = -w_m(w_mU_{xx} - U_{xl}) + (w_mU_{xl} - U_{ll}) = H^*/U_x^2 > 0$$

where $H^*$ is the bordered Hessian of the utility function $U$. Let $\Lambda_{ij}$ denote the cofactor of the $i$-$j$th element of $\Lambda$, e.g., $\Lambda_{11} = -(w_mU_{xx} - U_{xl})$. The slope of the moonlighting supply function with respect to $w_m$ is given by

$$(11) \qquad \frac{\partial L_m}{\partial w_m} = \frac{U_x}{\Lambda} - L_m\left(\frac{\Lambda_{11}}{\Lambda}\right)$$

This is of course the Hicks-Slutsky equation. The first term, $U_x/\Lambda$, which is unam-

[5] It can easily be shown that $\psi(w_m; Y_0)$, the indirect utility function, has the following partial derivatives:

$$\partial\psi/\partial w_m = -\lambda l \qquad \partial\psi/Y_0 = \lambda$$

[6] Let the first and second partials of $U$ be denoted in the conventional manner by subscripting $U$. The function $U$ is assumed to be twice continuously differentiable and quasi concave.

biguously positive, indicates that as the opportunity cost of *not* moonlighting increases, there is an increase in hours offered. The second term is the "income term." If leisure is a superior good, then $\Lambda_{11}/\Lambda$ is positive; the sign of the right-hand side of equation (11) is then ambiguous. If leisure is an inferior good, then $\Lambda_{11}/\Lambda$ is negative, and the income and substitution effects operate in the same direction.

The effect of an increase in the first-job wage rate $w_0$ can be seen from

$$\frac{\partial L_m}{\partial w_0} = - L_0\left(\frac{\Lambda_{11}}{\Lambda}\right) \tag{12}$$

which is negative if leisure is superior. The effect of a change in the number of first-job hours is given by

$$\frac{\partial L_m}{\partial L_0} = - 1 + \left(\frac{\Lambda_{11}}{\Lambda}\right)(w_m - w_0) \tag{13}$$

From equation (13), it can be seen that if leisure is superior and the moonlighting job pays less than the first job, then an increase in the number of hours contractually required on the first job will mean a decrease in the number of moonlighting hours offered. However, if the moonlighting job pays more than the first job, then the result is ambiguous. If both jobs pay the same, there is a one-for-one tradeoff in the supply function.

An increase in nonwage income can be expected to affect the supply of moonlighting labor. The magnitude of this effect is given simply by the income term, since

$$\frac{\partial L_m}{\partial A_0} = - \left(\frac{\Lambda_{11}}{\Lambda}\right) \tag{14}$$

## II. Data and Statistical Procedure

### A. *The Income Dynamics Panel*

A major problem in estimating the moonlighting supply curve has been the difficulty in finding an appropriate data base. Most data sources, for example the decennial Census, do not disaggregate yearly earnings between primary and secondary jobs. In contrast, the Income Dynamics Panel (*IDP*) of the University of Michigan's Survey Research Center does provide detailed information on family composition and earnings, as well as the wage rate and hours worked on the primary and secondary jobs of the family's head.[7]

The *IDP* contains a representative cross-section of the United States as well as a supplemental sample of families known to have low incomes. Members of the Panel interviewed in the spring of 1968, 1969, and 1970 supplied information on their employment experience in 1967, 1968, and 1969. The interviews were designed to collect information that explained short-term changes in the economic status of individuals and families. Between 1968 and 1970 the representative cross-section sample netted 2574 cases and the supplemental sample netted 1891 cases.

We divided the sample of male heads of households who were employed full time into two groups according to their participation in the secondary labor market. The first group consisted of 318 people who did some moonlighting in 1969. A second group of 1801 had not engaged in secondary employment in 1969. About 15 percent of those surveyed had second jobs—a somewhat greater percentage than may be expected in the general population. Table 1 presents the general profile of individuals in the two groups.

### B. *Statistical Procedure*

The average number of hours worked per week on the second job was made the

[7] For a complete discussion of this survey, see James Morgan and James Smith.

TABLE 1—SELECTED CHARACTERISTICS OF HOUSEHOLDS FROM THE INCOME DYNAMICS PANEL[a]

| | Moonlighters | Nonmoonlighters |
|---|---|---|
| **Head of Household** | | |
| Age, years | 37.2 (11.4) | 41.6 (12.2) |
| Percent nonwhite | 25.6 | 33.5 |
| Percent living in western United States | 16.8 | 14.3 |
| Percent living in urban areas | 45.3 | 46.0 |
| Percent high school graduates | 64.2 | 50.2 |
| Primary hourly wage rate | \$3.60 (\$1.83) | \$3.80 (\$2.57) |
| Weekly hours on primary job | 40.9 (9.4) | 43.6 (10.4) |
| Secondary hourly wage rate | \$3.40 (\$1.27) | 0.0 |
| Weekly hours on secondary job | 8.1 (9.1) | 0.0 |
| **Family** | | |
| Family size | 4.4 (2.3) | 4.2 (2.2) |
| Annual nonmoonlighting income | \$9,852 (\$5,109) | \$10,893 (\$7,355) |
| Annual asset income | \$ 361 (\$1,198) | \$ 516 (\$2,256) |
| Annual transfer payments | \$ 236 (\$702) | \$ 263 (\$905) |
| Annual labor income less moonlighting income | \$9,254 (\$4,828) | \$10,113 (\$6,666) |
| Annual cost of housing | \$1,235 (\$664) | \$ 1,146 (\$740) |
| **Sample Size** | 318 | 1,801 |

[a] Numbers in parentheses are standard deviations.

dependent variable in the statistical estimation of the moonlighting supply curve. This variable is characteristic of many economic variables obtained from surveys of households in that it has a limiting value and a large number of respondents grouped at that limit. As previously indicated, about 85 percent of the sample population were clustered at the limiting value of zero hours of moonlighting.[8]

As James Tobin has pointed out, "An explanatory variable in such a relationship may be expected to influence both the probability of limit responses and the size of nonlimit responses" (p. 24). If the probability of participation in the secondary job market were the only variable of interest, the probit model would be sufficient.[9] But using the probit model, Tobin points out, would be inefficient since information on the magnitude of the dependent variable (in this case, moonlighting hours) is being thrown away. Given a large concentration of observations at the zero moonlighting value, multiple regression analysis would not be appropriate either since the assumptions of that model would not be satisfied. In particular, a set of independent variables could take on values such that the expected value of the dependent variable, moonlighting hours, is below its limiting value 0. Clearly, this violates the assumption of a limited or bounded dependent variable.

The Tobit model, which combines the properties of the probit and multiple re-

[8] We could not assume that nonmoonlighters were in equilibrium at their first-job wage $w_0$ and hours $L_0$. Had such an assumption been made, then a moonlighting supply curve could have been estimated simply by *OLS* as follows:

for nonmoonlighters: $L_0 = g(w_0, A_0)$;

for moonlighters: $L_0 + L_m = g(w_m, A_0 + (w_0 - w_m)L_0)$

Essentially, by adjusting "nonlabor income" for moonlighters by $(w_0 - w_m)L_0$, total hours worked by both moonlighters and nonmoonlighters would be explained by the same equilibrium structure.

[9] In the probit model, the true proportion $p_i$ of heads of households with nonstochastic characteristics $x_{1i}$, $x_{2i}$, $x_{3i}$, . . . , $x_{ni}$ who moonlight would be estimated as $F(v_i)$ with

$$v_i = \beta_0 + \sum_{k=1}^{n} \beta_k x_{ki}$$

where $\beta_0$, $\beta_i$, . . . , $\beta_n$ are parameters to be estimated and $F$ is the cumulative normal distribution function.

gression models and is therefore more appropriate for an estimation of the moonlighting supply curve, selects a set of coefficients $\beta=(\beta_1, \ldots, \beta_n)$ that maximizes the limited dependent variable likelihood function. The first-order conditions for the maximum form a set of non-linear equations which can be computationally solved by an iterative procedure.

For the vector $X=(X_1, \ldots, X_n)$ consisting of observations on the set of independent variables, the unconditional expected value of the dependent variable $E(Y)$ is given by

$$E(Y) = P\cdot X\beta + (1-P)\cdot L + \sigma f\left(\frac{X\beta - L}{\sigma}\right) \tag{15}$$

where $L$ = limit value
$P$ = the probability of observing $Y>L$ given $X$, i.e., $P=\int_{-\infty}^{(X\beta-L)/\sigma} f(\xi)d\xi$
$\sigma$ = the standard error around the Tobit "index," $X\beta$
$f$ = ordinate of the unit normal density function[10]

The expected value of the dependent variable conditional on $Y>L$ is just $X\beta$.

## III. Estimating the Moonlighting Supply Function

One of the most important explanatory variables used in estimating the moonlighting supply function is the wage received or offered on the moonlighting job. For nonmoonlighters this information is not known. Therefore, we used the information on the 318 moonlighters to estimate a reduced form ($OLS$) equation with which we predicted the moonlighting wage for each member of the entire population. In effect, a two-stage procedure was performed. First, we projected moonlighting wages, based on a reduced form equation. Second, we fitted the data set of independent variables, including the reduced form prediction, to the Tobit model. This procedure is essentially an instrumental variable technique. Unfortunately, $OLS$ instrumental variable techniques lose their consistency property when combined with non-linear maximum likelihood procedures, such as the Tobit.[11]

Many factors could account for the variation in observed moonlighting wages. Typically in labor supply studies, variables reflecting personal characteristics such as education, race, age, and marital status are used along with variables reflecting regional labor market differences. We included these variables in our $OLS$ estimate of the moonlighting wage. In addition to these variables, we had another

[10] The elasticity of the predicted probability with respect to $X_i$ can be obtained as follows:

Let $$P = \int_{-\infty}^{(X\beta-L)/\sigma} f(\xi)d\xi = F\left(\frac{X\beta - L}{\sigma}\right)$$

then $$\frac{\partial P}{\partial X_i} = f\left(\frac{X\beta - L}{\sigma}\right)\frac{\beta_i}{\sigma}$$

and $$\frac{X_i}{P}\frac{\partial P}{\partial X_i} = \frac{\beta_i X_i}{\sigma}\frac{f\left(\frac{X\beta-L}{\sigma}\right)}{P} = \frac{\beta_i X_i}{\sigma}\frac{f\left(\frac{X\beta-L}{\sigma}\right)}{F\left(\frac{X\beta-L}{\sigma}\right)}$$

The elasticity of the expected value with respect to the variable $X_i$ can be obtained by differentiating equation (15).

$$\frac{\partial E(Y)}{\partial X_i} = P\beta_i$$

$$\frac{X_i}{E(Y)}\frac{\partial E(Y)}{\partial X_i} = \frac{\beta_i X_i F\left(\frac{X\beta - L}{\sigma}\right)}{E(Y)}$$

Further, if $z$ is a variable which appears in several $X_i$, then

$$\frac{\partial P}{\partial z} = \sum_{i=1}^{n}\frac{\partial P}{\partial X_i}\frac{\partial X_i}{\partial z}$$

$$\frac{\partial E(Y)}{\partial z} = \sum_{i=1}^{n}\frac{\partial E(Y)}{\partial X_i}\frac{\partial X_i}{\partial z}$$

[11] This point was made to us by T. Paul Schultz of Yale University. Consistent and asymptotically efficient estimators for this kind of problem have been developed. However, they are not as yet widely available.

valuable piece of information, namely the actual wage on the primary job. This variable can be thought of as reflecting a number of personal productivity characteristics not captured in the vector of demographic type variables. The *OLS* estimates for the reduced form of the moonlighting wage equation are:[12]

$$(16)\quad w_m = \underset{}{47.07} + \underset{(.062)}{.403}\, w_0 - \underset{(24.47)}{90.06}\, Race + \underset{(21.60)}{75.51}\, Urban + \underset{(23.42)}{47.33}\, Hisch + \underset{(27.62)}{113.64}\, Reg + \underset{(.94)}{2.26}\, Age$$

where

$$w_m = \text{moonlighting wage (¢/hr)}$$
$$w_0 = \text{primary wage (¢/hr)}$$
$$Race = \begin{cases} 0 & \text{white} \\ 1 & \text{nonwhite} \end{cases}$$
$$Urban = \begin{cases} 0 & \text{nonurban} \\ 1 & \text{urban} \end{cases}$$
$$Reg = \begin{cases} 0 & \text{non-West} \\ 1 & \text{West} \end{cases}$$
$$Hisch = \begin{cases} 0 & \text{nongraduate} \\ 1 & \text{high school graduate} \end{cases}$$
$$Age = \text{age (years)}$$
$$R^2 = .34$$
$$F(6,312) = 25.82$$

The above equation was used to predict the moonlighting wage rate $w_m$ for the entire sample population.[13]

To test the simple theory presented in Section I, equation (17) below was estimated using the Tobit procedure and the computed moonlighting wage rate from equation (16).

$$(17)\quad L_m = \alpha_0 + \alpha_1 w_m + \alpha_2 w_0 + \alpha_3 L_0 + \alpha_4 I_0 + \alpha_5 A_0 + \text{descriptive variables} + \text{error term}$$

A constraint on equation (17) is that $L_m \geq 0$ for each observation, where $w_m$ is the moonlighting wage rate, $w_0$ is the first-job wage rate, $L_0$ is first-job hours,[14] $I_0$ is an interaction term involving $w_m$, $w_0$, and $L_0$, and $A_0$ is a nonlabor income term. This form is useful because it allows for direct calculation of the partial derivatives of $L_m$ with respect to the variables $w_m$, $w_0$, $L_0$, and $A_0$.[15] Comparison of these partial

[12] Numbers in parentheses indicate the standard error of the coefficient. All of the variables in equation (16) had significant coefficients at the .05 level. Another regression with marital status included produced an insignificant coefficient for this variable. Ninety-two percent of the heads of households were married.

[13] This procedure may introduce a selectivity bias, which arises if moonlighting wage offers are stochastic *and* there is a systematic propensity for actually observed moonlighting wage rates to come from the upper (or lower) tail of the distribution. Imputing a moonlighting wage rate from observed wage rates usually overestimates the actual offers faced by those who do not moonlight. The effect of this overimputatation is to bias toward zero the effect of the moonlighting wage rate on the moonlighting hours in the Tobit equation. Selectivity biases are usually present in labor supply studies unless specific assumptions regarding the distribution of wage offers are made. We chose not to make such assumptions for this study. Instead, we simply note that the observed mean moonlighting wage was \$3.40 per hour and the estimated mean hourly moonlighting wage was \$3.41 for moonlighters and \$3.43 for nonmoonlighters. An alternative approach would be to use a totally reduced-form Tobit equation in which no estimate of the moonlighting wage offer appears, but this would thwart any effort to compare the moonlighting supply elasticity estimated from empirical work with the theoretical elasticity.

[14] $L_0$ is considered an independent variable which, according to the simple theory, is fixed for each individual; across individuals it is a parameter which can take on many values. The question of when $L_0$ is "fixed" (before or after he chooses his first job) is critical because if he is able to choose from a set of first jobs which offers a range of working hours, then utility maximization would require that the best combination of $L_0$ and $L_m$ be chosen. Our theory, however, views $L_0$ as fixed and not subject to manipulation by the individual. We observe that a large proportion of full time jobs generally demand some fixed commitment around 35 to 40 hours per week and that this artifact is an important feature of the *U.S.* labor market.

[15] Equation (17) as a Tobit index can also be viewed in the following instructive way: moonlighting, according to the theory, occurs when the moonlighting wage offer $w_m$ exceeds the second-job reservation wage $w_m^r$. The reservation wage for an individual may plausibly be made a function of $w_0$, $L_0$, and $A_0$, which formally

TABLE 2—COMPARISON OF SIMPLE THEORY AND EQUATION (17)[a]

| Partial Derivative | From simple theory | From equation (17): Specification $A$ | Specification $B$ |
|---|---|---|---|
| $\frac{\partial L_m}{\partial w_m}$ | $\frac{U_x}{\Lambda} - L_m\left[\frac{\Lambda_{11}}{\Lambda}\right]$ | $\alpha_1 + \alpha_4 L_0$ | $\beta_1$ |
| $\frac{\partial L_m}{\partial w_0}$ | $-L_0\left[\frac{\Lambda_{11}}{\Lambda}\right]$ | $\alpha_2 - \alpha_4 L_0$ | $\beta_2 + \beta_4 L_0$ |
| $\frac{\partial L_m}{\partial L_0}$ | $-1 + \left[\frac{\Lambda_{11}}{\Lambda}\right](w_m - w_0)$ | $\alpha_3 + \alpha_4(w_m - w_0)$ | $\beta_3 + \beta_4 w_0$ |
| $\frac{\partial L_m}{\partial A_0}$ | $-\left[\frac{\Lambda_{11}}{\Lambda}\right]$ | $\alpha_5$ | $\beta_5$ |

[a] For purposes of exposition, and to distinguish the two models, let the coefficients of equation (17) be $\alpha$ for specification of $A$ and $\beta$ for specification of $B$.

derivatives with equations (11) through (14) provides some testable hypotheses of the simple theory.

Equation (17) was actually estimated separately for two alternative specifications: in specification $A$, the interaction term $I_0$ was $(w_m - w_0)L_0$; in specification $B$ the interaction term was $w_0 L_0 + Z_0$, where $Z_0$ is labor income earned by members of the family other than the head of the household, so that $I_0$ was in effect total family labor income less moonlighting income. We chose to use alternative specifications because the different interaction terms yield different implications about the moonlighting supply curve. In specification $A$, the pure substitution term (in the neighborhood of equilibrium) is a linear function of $L_m$ and $L_0$, while in specification $B$ the pure substitution term is a linear function of $L_m$ only. Table 2 establishes the crucial links between the theoretical model and the empirical work.

Table 3 presents the results of fitting the *IDP* data and the computed moonlighting wage, equation (16), to both specifications of equation (17) using the Tobit procedure.

accounts for their presence in equation (17). If $\Phi$ is the cumulative distribution function for the second-job reservation wage $w_m^r$, then $\Phi(w_m) = P(w_m^r \leq w_m)$ = the probability of moonlighting given a moonlighting offer of $w_m = F$ (the normalized Tobit index).

Unlike many labor supply studies, the estimated effect on the supply of secondary labor to changes in the wage, primary hours, and nonlabor income variables is remarkably insensitive to the particular specification, i.e., $A$ or $B$. Note that the coefficients of $w_m$ and $w_0$ are statistically insignificant in specification $A$ while they are significant in specification $B$. Yet as Table 4 shows, the estimated changes in $L_m$ due to marginal changes in the variables $w_m$, $w_0$, $L_0$, and $A_0$ (evaluated at the mean value of the independent variables taking just moonlighters) are not statistically different.

## IV. Test of the Simple Theory

The simple theory was based on an individual for whom $L_m > 0$; for this individual the Tobit index is the moonlighting supply curve. The simple theory can be tested against the data by noting the relationships among the estimated coefficients of equation (17), as presented in Table 2.

For each specification a hypothesis about each partial derivative was constructed by setting the expression for that partial derivative from equation (17) equal

TABLE 3—TOBIT REGRESSION RESULTS
(Dependent variable, weekly hours on second job)

| Independent Variable | Coefficient | Asymptotically normal $z$-score[a] | Elasticity of expected value[b] | Elasticity of predicted probability[b] |
|---|---|---|---|---|
| **Specification A** | | | | |
| Secondary wage rate, $w_m$($/hr) | 1.109 | .649 | .447 | .363 |
| Primary wage rate, $w_0$($/hr) | −.2843 | −.180 | −.126 | −.102 |
| Weekly hours on first job, $L_0$ | −.2768 | −4.588 | −1.406 | −1.143 |
| Interaction term, $I_0$($/wk)[c] | .0458 | 1.264 | 0.074 | −.060 |
| Nonlabor income, $A_0$($/wk) | −.0321 | −.839 | −.019 | −.015 |
| Age (years) | −.2907 | −5.954 | −1.400 | −1.138 |
| Family size | .7046 | 2.822 | .350 | .284 |
| Constant | .4191 | | | |
| **Specification B** | | | | |
| Secondary wage rate, $w_m$($/hr) | 3.153 | 4.114 | 1.270 | 1.031 |
| Primary wage rate, $w_0$($/hr) | −1.944 | −3.166 | −.862 | −.700 |
| Weekly hours on first job, $L_0$ | −.2472 | −3.849 | −1.255 | −1.020 |
| Interaction term, $I_0$($/wk)[d] | −.00776 | −.835 | −.175 | −.142 |
| Nonlabor income, $A_0$($/wk) | −.0344 | −.892 | −.020 | −.016 |
| Age (years) | −.2910 | | −1.401 | −1.138 |
| Family Size | .7147 | | .355 | .288 |
| Constant | −.7191 | | | |

*Note:*
Predicted probability evaluated at mean $X(I)$: .1353 for Specification $A$ and .1357 for Specification $B$.
Observed probability, .1501; Mean dependent variable, 1.21
Expected value, $E(y)$, evaluated at mean $X(I)$: 1.150 for Specification $A$ and 1.154 for Specification $B$.
Estimated standard error around Tobit index, 16.82
Equation $\chi^2(7)$: 188.14 for Specification $A$ and 187.34 for Specification $B$.

[a] In Tobit, $z$-scores are used to test for coefficient significance and are computed from the coefficients and their standard errors just as $t$-statistics are. In large samples, $z$-scores are indistinguishable from corresponding $t$-statistics.

[b] The elasticity refers to just the independent variable listed on the left holding the other variables constant, and is evaluated at mean values of all the independent variables.

[c] $I_0=(w_m-w_0)L_0$

[d] $I_0=w_0L_0+Z_0$

to the expression derived from the simple theory and substituting the appropriate estimate of $\Lambda_{11}/\Lambda$, i.e., $\alpha_5$ or $\beta_5$. The first null hypothesis which relates to $\partial L_m/\partial w_m$ is a test of whether the pure substitution term as estimated is positive. The resulting $t$-statistics are large enough to confirm the positiveness of the substitution term for both specifications.[16] See Table 5.

Hypotheses (2) and (3) derive from the partial derivatives $\partial L_m/\partial w_0$ and $\partial L_m/\partial L_0$,

[16] For specification $A$ the substitution term was 3.242 (hrs/wk)/($/hr) and for specification $B$ it was 3.432 (hrs/wk)/($/hr), which again are quantitatively similar (at the means of the independent variables) despite the different econometric specifications.

TABLE 4—ESTIMATED RESPONSE OF $L_m$
(Moonlighters only)

| Partial Derivative | Specification $A$[a] | Specification $B$[a] |
|---|---|---|
| $\dfrac{\partial L_m}{\partial w_m}$ | 2.982 ($t$ = 3.922) | 3.153 ($t$ = 4.114) |
| $\dfrac{\partial L_m}{\partial w_0}$ | −1.589 ($t$ = −3.024) | −2.2616 ($t$ = −4.249) |
| $\dfrac{\partial L_m}{\partial L_0}$ | −.2857 ($t$ = −4.627) | −.2752 ($t$ = −4.539) |
| $\dfrac{\partial L_m}{\partial A_0}$ | −.03212 ($t$ = −.839) | −.0344 ($t$ = −.892) |

[a] Evaluated at the mean value of the independent variables using nonlimit observations.

TABLE 5—TESTS OF SIMPLE THEORY[a]

| Hypothesis | Specification $A$ | $t$-statistic | Specification $B$ | $t$-statistic |
|---|---|---|---|---|
| (1) | $H_0$: $\hat{\alpha}_1+\hat{\alpha}_4\bar{L}_0-\hat{\alpha}_5\bar{L}_m>0$ | 3.792 | $H_0$: $\hat{\beta}_1-\hat{\beta}_5\bar{L}_m>0$ | 3.971 |
| (2) | $H_0$: $\hat{\alpha}_2-(\hat{\alpha}_4+\hat{\alpha}_5)\bar{L}_0=0$ | .534 | $H_0$: $\hat{\beta}_2+(\hat{\beta}_4-\hat{\beta}_5)\bar{L}_0=0$ | .541 |
| (3) | $H_0$: $\hat{\alpha}_3+(\hat{\alpha}_4+\hat{\alpha}_5)(\bar{w}_m-\bar{w}_0)+1=0$ | −11.167 | $H_0$: $\hat{\beta}_3+\hat{\beta}_4\bar{w}_0+\hat{\beta}_5(\bar{w}_m-\bar{w}_0)+1=0$ | −12.216 |

[a] Variables with bars over them indicate that they are to be evaluated at the mean using the nonlimit (i.e., moonlighters) subsample.

respectively. For both specifications one cannot reject the second null hypothesis. Indeed, it seems reasonable to say that the simple theory is confirmed here. However, for both specifications, one must reject the third hypothesis, and this represents a a failure of the simple theory. This, of course, does not mean that the simple theory is to be discarded.

## V. Some Implications

As mentioned earlier, a change in an explanatory variable in a Tobit regression is likely to affect the probability of a limit response and the magnitude of nonlimit responses. How might the total supply of moonlighting labor be affected by a change in $w_m$, $w_0$, and $L_0$? The results presented in Table 3 indicate that the supply of moonlighting labor increases with the moonlighting wage rate $w_m$ and falls with primary job earnings $w_0L_0$. An increase in the moonlighting wage rate will increase the labor supplied by moonlighters and cause previous nonmoonlighters to enter the secondary market. To put this in quantitative terms, specification $A$ would predict that a 10 percent increase in the moonlighting wage rate results in moonlighters increasing their hours worked by 26.0 percent. This can be obtained easily by converting the slope in Table 4, which is conditional on $L_m>0$, to an elasticity. More importantly, unconditional expected moonlighting hours worked increase 17.7 percent while the probability of entering the secondary labor market increases by 9.6 percent.

A given increase in first-job earnings will have a negative effect, as seen by the signs on the primary wage, the primary hours, and interaction terms. However, the magnitude of the change depends upon whether the increase was affected by a change in the primary wage rate or by a change in hours worked on the primary job. The negative elasticity with respect to primary hours appears to be greater in absolute value than the negative elasticity with respect to primary wage, because a change in the latter affects only earnings but a change in the former also reduces the time available to moonlight.[17]

Family size (a proxy for consumption) is significant and positively related to moonlighting hours. Furthermore, as is consistent with the life cycle consumption

[17] There may be a certain amount of simultaneity in the determination of the first-job wage rate and first-job hours: in fact as stated earlier $w_0$ and $L_0$ may be negotiated together. In addition, we have not taken income tax liabilities into account since both $w_0$ and $w_m$ are pretax wage rates. If the avoidance of income taxes is systematically easier or more prevalent for moonlighting income than for primary income we have understated the relative return to moonlighting. If there are fixed costs associated with a moonlighting job, then the simple theory would predict a higher reservation wage and a discontinuity in an individual's moonlighting supply curve at the new reservation wage. Further, if moonlighting does occur and leisure is superior, the number of moonlighting hours offered at each such $w_m$ is greater if there are fixed costs than if there are no fixed costs. Unfortunately, it was not possible to incorporate fixed-job costs into our empirical work with the available data.

hypothesis, age which can be considered an inverse proxy for unmet family needs shows a significantly negative relationship to moonlighting hours.

REFERENCES

A. Grossman, *Multiple Jobholding in May 1974*, Special Labor Report 177, U.S. Department of Labor.

H. W. Guthrie, "Who Moonlights and Why," *Ill. Bus. Rev.*, Mar. 1965, *22*, 6–8.

———, "Teachers in the Moonlight," *Mon. Lab. Rev.*, Feb. 1969, *92*, 28–31.

H. R. Hamel, "Moonlighting—An Economic Phenomenon," *Mon. Lab. Rev.*, Oct. 1967, *90*, 17–22.

L. N. Moses, "Income, Leisure, and Wage Pressure," *Econ. J.*, June 1962, *72*, 320–34.

J. N. Morgan and J. D. Smith, *A Panel Study of Income Dynamics*, Vols. I-III, Inst. Soc. Res., Survey Res. Center, Univ. Michigan, Ann Arbor 1969.

R. Perlman, *Labor Theory*, New York 1969.

J. Tobin, "Estimation of Relationships for Limited Dependent Variables," *Econometrica*, Jan. 1958, *26*, 24–36.

# Interdependent Preferences

By Robert A. Pollak*

In this paper I set out a model of interdependent preferences—preferences which depend on other people's consumption—and examine its implications for demand behavior. It is a commonplace that preferences are influenced by other people's consumption, but this insight has never been incorporated into demand analysis in a satisfactory way. I believe that tractable models of interdependent preferences can be constructed in much the same way that the other major form of endogenous taste change, habit formation, was analyzed in my 1970 article; indeed, I suggested in that paper that the general approach used there could also be used to analyze interdependent preferences.

In Section I, I specify a short-run model of interdependent preferences, "linear" interdependence. I introduce linear interdependence in the context of the linear expenditure system by postulating that the necessary quantities depend linearly on other people's past consumption. By specifying that interdependent preferences operate through *past* consumption, I avoid the need to determine everyone's consumption simultaneously, and by working with a complete system of demand functions, I guarantee that the budget constraint will be satisfied. I consider a number of alternative specifications of the interdependence relationship, and extend the results from the linear expenditure system to more general systems of demand functions. The discussion in Section I focuses on individual behavior in the short run.

In Section II, I investigate the per capita demand functions implied by the model of interdependent preferences introduced in Section I; in Section III, the long-run or steady-state equilibrium of these per capita demand functions is characterized. I emphasize the per capita rather than the individual demand functions because the outcome of the mutual interaction of consumption and tastes which characterized interdependent preferences is manifested only in the long-run per capita demand functions.

Finally, in Section IV, I summarize the discussion of interdependent preferences, discuss simultaneous interdependence, a specification in which preferences are influenced by other people's current consumption, suggest several areas in which further work is needed, and discuss a problem which interdependent preferences poses for welfare economics.

One of the most interesting conclusions to emerge from this examination of interdependent preferences has to do with the role of expenditure distribution as a determinant of per capita consumption patterns.[1] One might expect that with interdependent preferences expenditure distribution would play a significant role. But analysis of a number of different specifications shows that this is not the case; with some it matters, while with others it does not.

Another important conclusion is that many specifications of interdependent

* University of Pennsylvania. This research was supported in part by the National Science Foundation. I am grateful to an anonymous referee for helpful comments.

[1] Throughout this paper I have refrained from using "income" to mean "total expenditure on goods and services"; instead, I have used "expenditure." The terminology takes some getting used to, especially when discussing the effect of the "distribution of expenditure" on the per capita consumption pattern.

preferences cannot be distinguished from linear habit formation on the basis of the implied per capita demand behavior. In these cases it is only by observing individual demand behavior in successive periods that we can distinguish between interdependent preferences and habit formation.

The literature on interdependent preferences in demand analysis is relatively sparse. The lead provided by James Duesenberry was never systematically explored. S. J. Prais and H. S. Houthakker (Section 2.5) discuss a model of interdependent preference in a budget study context, but their analysis refers to a single commodity and is never integrated into a theory of maximizing behavior. An interesting paper by Wulf Gaertner develops a model of interdependent preferences which parallels some of the analysis presented here.

## I. Individual Demand Functions in the Short Run

In this section I examine the individual demand functions corresponding to a number of alternative specifications of interdependent preferences. For expositional convenience, they are introduced and discussed in the context of the linear expenditure system; it is then argued that they can be generalized to a wide variety of systems of demand functions.

In the linear expenditure system the demand function of individual $r$ for good $i$ in period t is given by

$$(1) \quad \overset{r}{q}_{it} = h^{rit}(P_t, \mu_t)$$
$$= \overset{r}{b}_{it} - \frac{\overset{r}{a}_i}{p_{it}} \sum_k p_{kt}\overset{r}{b}_{kt} + \frac{\overset{r}{a}_i}{p_{it}} \overset{r}{\mu}_t$$

Individuals are identified by superscripts on $a$, $b$, $q$, and $\mu$, but everyone faces the same prices.

I propose to incorporate interdependent preferences into the linear expenditure system by postulating that some of the parameters of the demand functions depend on other people's consumption. For reasons of tractability, I shall assume that the $a$ are constant and that interdependence operates through the $b$. The most straightforward assumption is that the $b$ depend linearly on other people's consumption:

$$(2) \qquad \overset{r}{b}_{it} = \overset{r*}{b}_i + \sum_{\substack{s=1 \\ s \neq r}}^{R} \overset{rs}{\beta}_i \overset{s}{q}_{it}$$

where $R$ is the number of individuals.[2] As a specification of interdependent preferences, (2) has two major defects. First, it involves too many interdependence coefficients—$n \times R-1$ for each individual. And second, because each person's tastes depend on everyone else's current consumption, the determination of an equilibrium consumption pattern for everyone in the society is a formidable task.

Instead of specifying that each individual's preferences depend on everyone else's current consumption, I shall assume that they depend on other people's past consumption. This assumption has the merit of analytical tractability, a virtue not to be despised. It is also consistent with the belief that the acquisition of preferences is part of the process of socialization. It is tempting to argue that lagged interdependence is more plausible than simultaneous interdependence, but since nothing has been said about the length of the periods, such arguments are fragile.

To incorporate lagged interdependence into the linear expenditure system, I postulate that $\overset{r}{b}_{it}$ depends linearly on each individual's consumption of good $i$ in the previous period:

[2] One would expect the $\beta$ to be nonnegative, so that an increase in someone else's consumption of a good implies, *ceteris paribus*, an increase in its consumption by individual $r$. As in habit formation models, dynamic stability conditions imply restrictions on the $\beta$; these are discussed briefly in Section IV.

$$(3)\qquad b_{it}^{r} = b_i^{r*} + \sum_{s=1}^{R} \beta_i^{rs} q_{it-1}^{s}$$

This specification differs from (2) both because it embodies the hypothesis of *lagged* rather than simultaneous interdependence, and because the summation in (3) runs over *all* individuals, not just *all other* individuals. One might argue that pure interdependence requires the "own effect" $\beta_i^{rr}$ to be zero, and that when it is nonzero the interdependence model is contaminated by traces of habit formation, but I prefer the more general model, especially since we are free to consider pure interdependence as a special case.

The linear interdependence specification (3) involves an unmanageably large number of interdependence coefficients or "weights": unless there is a systematic relationship among the $\beta$, each individual assigns $n \times R$ independent weights. In order to reduce the number of interdependence parameters to manageable proportions, I now consider alternative assumptions about the $\beta$.

I begin by treating two closely related cases: (a), in the first, individual $r$ gives equal weight to everyone's past consumption, including his own:

$$(4a)\qquad \beta_i^{rs} = \beta_i^{r} \qquad s = 1, \ldots, R$$

and, (b), in the second, he gives equal weight to everyone else's past consumption and 0 weight to his own

$$(4b)\qquad \beta_i^{rs} = \begin{cases} \beta_i^{r} & s=1,\ldots,r-1, r+1,\ldots,R \\ 0 & s=r \end{cases}$$

In both of these cases it is possible to discuss interdependent preferences in terms of *per capita* past consumption, $\bar{q}_{it-1}$.

Under (4a), (3) becomes

$$(5a)\qquad b_{it}^{r} = b_i^{r*} + \hat{\beta}_i^{r} \bar{q}_{it-1}$$

where

$$(6)\qquad \bar{q}_{it-1} = \frac{1}{R} \sum_{s=1}^{R} q_{it-1}^{s} \quad \text{and} \quad \hat{\beta}_i^{r} = R\beta_i^{r}$$

Under (4b), it becomes

$$(5b)\qquad b_{it}^{r} = b_i^{r*} + \hat{\beta}_i^{r} \bar{q}_{it-1} - \beta_i^{r} q_{it-1}^{r}$$

In both cases, $b_{it}^{r}$ depends linearly on per capita past consumption, although in (5b) an adjustment is made for the individual's own past consumption.

Substituting (5a) or (5b) into the linear expenditure system (1) and dropping individual subscripts on $a$ and time subscripts on $p$ and $\mu$ yields the demand functions of individual $r$:[3]

$$(7a)\qquad q_{it}^{r} = b_i^{r*} - \frac{a_i}{p_i} \sum p_k b_k^{r*} + \frac{a_i}{p_i} \mu + \hat{\beta}_i^{r} \bar{q}_{it-1} - \frac{a_i}{p_i} \sum p_k \hat{\beta}_k^{r} \bar{q}_{kt-1}$$

or

$$(7b)\qquad q_{it}^{r} = b_i^{r*} - \frac{a_i}{p_i} \sum p_k b_k^{r*} + \frac{a_i}{p_i} \mu + \hat{\beta}_i^{r} \bar{q}_{it-1} - \frac{a_i}{p_i} \sum p_k \hat{\beta}_k^{r} \bar{q}_{kt-1} - \beta_i^{r} q_{it-1}^{r} + \frac{a_i}{p_i} \sum p_k \beta_k^{r} q_{kt-1}^{r}$$

Duesenberry, p. 27, points out that changes in tastes are caused by frequent contact with superior goods, not by mere knowledge of their existence. That is, a

[3] This model of interdependent preferences, like the habit model, can be extended in a number of directions. For example, per capita consumption in the more distant past may influence current tastes or $b$ may depend linearly on the previous peak of per capita consumption, with or without a "discount factor" which causes the influence of the peak to diminish as it recedes into the past. This introduces a "ratchet effect" of the type proposed by Duesenberry, p. 115. The linearity assumption can be dropped to permit $b$ to be non-linear functions of per capita consumption in the previous period, or, more generally, of per capita consumption in the more distant past.

man's preferences will be influenced more by the consumption of those with whom he has close contact than by those with whom his contact is more distant. Duesenberry's observation implies that individual demand depends on the distribution of consumption in the previous period, not just on its average; if he is correct, then the assumption that the $\beta$ are the same for all individuals is unsatisfactory and individual consumption patterns will depend on the distribution of consumption in the previous period.

I now describe several models of interdependent preferences based on the premise that individuals are arrayed in a hierarchy in which each individual's preferences are influenced by the consumption behavior of those above him. For definiteness, one can think of each individual as attempting to imitate the consumption behavior of individuals who he perceives as having higher status, but this interpretation is not necessary. Position in the hierarchy is defined in terms of who influences whom, and no sociological concepts need be invoked.[4] It is convenient to number individuals in terms of their position in the hierarchy, so that Mr. 1 is at the top and Mr. $R$ at the bottom.

A simple model of interdependent preferences can be built on the assumption that each individual is concerned only with the consumption of the man one rung above him in the hierarchy.[5] To close the model the behavior of the first man must be specified. Since there is no one above him, it could be argued that the most appropriate specification is $\beta_i^{1s}=0$ for all $s$, so that $b_{it}^1=b_i^{1*}$. This, however, implies a sharp break between the preference pattern of the first man and those immediately below him. I prefer to assume that Mr. 1's preferences depend on his own past consumption, as in the habit formation model.

A "two-class" model of interdependent preferences in which the members of the lower class ($L$) emulate the consumption standards of the upper class ($U$) is another version of the hierarchy model. Again working with the linear expenditure system, suppose that the necessary basket of an individual in the lower class depends linearly on the average consumption of those in the upper class. We could assume that for members of the upper class $b_{it}^r=b_i^{r*}$, but I prefer to postulate that their tastes are influenced by the average past consumption of their own class.

Any of the specifications of interdependent preferences can be applied to any system of demand functions locally linear in total expenditure.[6] W. M. Gorman has shown that any system of demand functions locally linear in total expenditure and derivable from a well-behaved utility function are of the form

$$(8)\quad q_{it}^r = b_i + f_i - \gamma^i \sum p_k b_k - \gamma^i f + \gamma^i \mu$$

where $\gamma^i=g_i/g$ and $f(p)$ and $g(p)$ are functions homogeneous of degree one. To introduce interdependent preferences, we postulate that the $b$ are given by one of the specifications of interdependent preferences discussed above.

More generally, any of these specifications of interdependent preferences can be applied to systems of demand functions which are non-linear in income, provided the system contains $n$ independent parameters, $b_t=(b_{1t}, \ldots, b_{nt})$: $q_t=h(P_t, \mu_t; b_t)$, but the resulting system may be nonlinear in past consumption.

[4] To simplify the exposition, I assume that there are no "ties," although these could be accommodated without altering the substance of the analysis.

[5] An individual need not recognize the entire hierarchy but only the man above him. The hypothesis is that patterns of influence are consistent with an underlying structure in which individuals are arrayed in a one-dimensional hierarchy.

[6] This is not quite true; regularity conditions may cause trouble.

## II. Per Capita Demand Functions in the Short Run

In Section I, I examined a number of alternative specifications of "linear" interdependent preferences and discussed their implications for the short-run demand behavior of an individual. In this section, I consider a society made up of individuals whose preferences are interdependent and examine the implied short-run per capita demand functions. The emphasis on per capita rather than market demand functions is a matter of form rather than substance. It is convenient to state results in this way to facilitate comparisons with both the individual demand functions of Section I and the habit formation model of my 1970 article.

Although the short-run per capita demand functions are of interest for their own sake, I treat them primarily as a bridge between the short-run individual demand functions of Section I and the long-run per capita demand functions of Section III. I emphasize the long-run per capita demand functions because the ramifications of interdependent preferences are only apparent in the long run, when the full consequences of each person's consumption on everyone else's preferences have manifested themselves.

I base the exposition on the linear expenditure system, but the results generalize immediately to any system of demand functions locally linear in total expenditure. As was pointed out in Section I, the specifications of interdependent preferences introduced there can be applied to individual demand functions which are non-linear in total expenditure. But even in the absence of interdependent preferences, the per capita demand functions implied by systems of individual demand functions non-linear in total expenditure depend on the distribution of expenditure among individuals. In Section III, when I discuss the long-run per capita demand functions, I shall focus on the effect of the distribution of expenditure among individuals on the per capita consumption pattern. In order to isolate the effect of interdependent preferences, I consider only short-run individual demand functions which are locally linear in total expenditure. In this case, provided everyone has the same marginal budget shares, if the distribution of expenditure has any influence on the long-run per capita consumption pattern, it can be attributed to interdependent preferences.

If the short-run per capita demand functions are locally linear in per capita expenditure, then the distribution of expenditure has no direct effect on the long-run consumption pattern; any influence it has must be indirect and operate through interdependent preferences. If the short-run per capita demand functions are locally linear in per capita expenditure and independent of the distribution of past consumption, then the long-run demand functions will be independent of the distribution of per capita expenditure. Under some specifications of interdependent preferences, the distribution of past consumption is a significant determinant of the short-run consumption pattern, while under others it is not. In this section I examine the short-run per capita demand functions corresponding to various specifications of interdependent preferences and see which ones fall into each category.

I opened the discussion of interdependent preferences in Section I by considering two specifications in which an individual gives equal weight to everyone else's past consumption. They differed in that one specification, (4a), gives the individual's own past consumption the same weight as anyone else's, while the other, (4b), gives it 0 weight. With the linear expenditure system, the individual demand functions corresponding to these two specifications are given by (7a) and (7b), re-

spectively. To examine the market or per capita demand functions implied by these individual demand functions, we assume that everyone has the same marginal budget shares and the same interdependence coefficients; that is,

$$(9) \qquad a_i^r = a_i \qquad r = 1, \ldots, R$$

$$(10) \qquad \beta_i^r = \beta_i \qquad r = 1, \ldots, R$$

These assumptions enable us to drop the individual superscripts on the $a$ and $\beta$. Individual differences in demand behavior reflect differences in total expenditure or differences in tastes, but the only admissible differences in tastes are differences in the $b^*$.

The per capita demand functions are obtained by summing the individual demand functions over all individuals and dividing by $R$. If each individual gives equal weight to everyone's past consumption including his own, (4a), this yields

$$(11a) \quad \bar{q}_{it} = \bar{b}_i^* - \frac{a_i}{p_i} \sum p_k \bar{b}_k^* + \frac{a_i}{p_i} \bar{\mu}_t + \hat{\beta}_i \bar{q}_{it-1} - \frac{a_i}{p_i} \sum p_k \hat{\beta}_k \bar{q}_{kt-1}$$

where $\bar{b}^*$ and $\bar{\mu}_t$ denote the average values of $b_i^{r*}$ and $\mu_t^r$:

$$(12) \qquad \bar{b}_i^* = \frac{1}{R} \sum_{s=1}^{R} b_i^{s*} \qquad \bar{\mu}_t = \frac{1}{R} \sum_{s=1}^{R} \mu_t^s$$

If each individual gives equal weight to everyone else's consumption in the previous period and 0 weight to his own, summing the individual demand functions (7b) and dividing by $R$ yields

$$(11b) \quad \bar{q}_{it} = \bar{b}_i^* - \frac{a_i}{p_i} \sum p_k \bar{b}_k^* + \frac{a_i}{p_i} \bar{\mu}_t + \overset{*}{\beta}_i \bar{q}_{it-1} - \frac{a_i}{p_i} \sum p_k \overset{*}{\beta}_k \bar{q}_{kt-1}$$

where $\beta_i^*$ is defined by

$$(13) \qquad \overset{*}{\beta}_i = (R-1)\beta_i = \hat{\beta}_i - \beta_i$$

Thus, the alternative specifications (4a) and (4b) lead to per capita demand functions which are essentially identical. It would not be surprising to find that they imply similar per capita demand functions —if the number of individuals is large and income is fairly evenly distributed, the difference between the average of everyone's past consumption and the average of everyone else's past consumption must be small. But my assertion is stronger; the per capita demand functions (11a) and (11b) differ only in that the interpretation of one coefficient in terms of individual demand functions is slightly different in the two cases. The systems of equations to be estimated from market data are indistinguishable.

In both of these cases the distribution of consumption in the previous period has no influence on the current per capita consumption pattern. In (11a), when each individual is influenced by the average of everyone's past consumption including his own, this is trivially obvious: since the individual demand functions are independent of the distribution of past consumption, so are the per capita demand functions. In (11b), when each individual is influenced by the average of everyone else's past consumption, individual demand is influenced to some degree by the distribution of past consumption, although the influence is small unless the community is small or the distribution of past consumption very unequal. In the limit, as the community grows large or as the distribution of past consumption approaches equality, the individual demand functions approach independence of the distribution of past consumption.

My claim about the per capita demand functions is not merely an assertion about these limiting cases. It does not depend on the size of the community or the degree of

inequality in the distribution of past consumption. When each individual is influenced by the average of everyone else's past consumption, the per capita demand functions depend only on average past consumption, and not on its distribution.

I now examine the implications for the per capita demand functions of the models of interdependent preferences based on a hierarchy in which each individual is influenced by the consumption behavior of those above him. Again assuming that everyone has the same $a$ and $\beta$, (9) and (10), the market demand functions, are obtained by summing over all individuals. When each individual is concerned with the consumption of the man above him in the hierarchy, one might expect the distribution of past consumption to be a major determinant of the present per capita consumption pattern. Surprisingly enough, it is not. The only parameter of the distribution of past consumption other than its average which enters the per capita demand functions is the range of the distribution, $q^1_{it-1}-q^R_{it-1}$.[7] But the range enters only in terms of the form $\bar{q}_{it-1}+(q^1_{it-1}-q^R_{it-1})/R$, and average past consumption, $\bar{q}_{it-1}$, is likely to swamp the effects of the range, since the latter is divided by $R$. Unless the community is small or the distribution of past consumption very unequal, the influence of the distribution of past consumption is small.

In the hierarchy model, the consumption pattern of each individual depends on the past consumption of the man above him, but not on anyone else's past consumption. Hence, the distribution of past consumption has a substantial impact on individual consumption patterns. But while individual demand functions depend significantly on the distribution of past consumption, the per capita demand functions do not. There is nothing paradoxical about this. In the habit models of my 1970 article, each individual's demand functions depend only on his past consumption, yet it is easy to verify that the implied per capita demand functions are independent of the distribution of past consumption. Similarly, with interdependent preferences, if each individual is concerned with the average of everyone else's past consumption, the individual demand functions depend on the distribution of past consumption but the per capita demand functions do not. In general, if the individual demand functions are independent of the distribution of past consumption, then so are the per capita demand functions. But the converse need not hold: as this example shows, the per capita demand functions may be independent of the distribution of past consumption even when the individual demand functions are not.

In the two-class models the current per capita consumption pattern does depend significantly on the distribution of past consumption. The model assumes that everyone has the same marginal budget shares, (9), and the same interdependence coefficients, (10); since the lower class's $\beta$ represent the influence of higher status individuals while the upper class's $\beta$ represent the influence of their peers, this is a very strong assumption.

### III. Long-Run Demand Functions

In this section I examine the "long-run" per capita demand functions corresponding to various specifications of interdependent preferences. The definition of the long-run or steady-state demand functions in the interdependent preferences model is essentially the same as in the habit model (see my 1970 article, p. 751). Given the per capita consumption vector of period one, the short-run per capita demand functions yield a consumption vector for period one.

[7] Strictly speaking, this is not the range of the distribution since we cannot identify $q^1$ and $q^R$ with the highest and lowest consumption in the community.

In a "steady-state" or "long-run equilibrium" the consumption vector in every subsequent period will also be equal to the consumption vector of period zero. The long-run demand functions associate a steady-state consumption pattern with every price-income situation.[8]

It is only in the long run that one can observe the full effects of interdependent preferences. In the short run, the distribution of expenditure among individuals can influence the per capita consumption pattern only if different individuals have different marginal budget shares. When the short-run individual demand functions are locally linear in expenditure and everyone has the same marginal budget shares, then the distribution of expenditure among individuals has no direct influence on the short-run per capita consumption pattern, but the distribution of expenditure can influence the long-run consumption pattern indirectly, through the interdependent preferences.

The exposition in Section II was based on the linear expenditure system. In this section, I use a more general framework, namely, demand functions locally linear in expenditure. I emphasized in Section II that the results obtained there extend immediately to this more general setting. I use it here to emphasize the similarity of the long-run per capita demand functions which arise in the model of interdependent preferences with the long-run individual demand functions of the habit formation model of my articles cited in the references.

I begin by discussing specifications of interdependent preferences in which each individual is influenced by average past consumption. There are two such specifications, (4a) and (4b), which differ only in their treatment of the individual's own past consumption. For definiteness, I shall discuss the specification in which each individual is influenced by the average of everyone else's past consumption, but since both specifications imply the same short-run per capita demand functions, they must also imply the same long-run per capita demand functions.

[8] This assumes the existence of such a steady state.

The long-run per capita demand functions can be found by solving this system of short-run demand functions under the assumption that $\bar{q}_{it}=\bar{q}_{it-1}=\bar{q}_i$, for all $i$. Mathematically, this problem is identical to one considered in my forthcoming article where it was found necessary to solve the system of short-run demand functions implied by habit formation to find the long-run demand functions. There is no need to repeat here the argument used there; I shall simply assert its conclusion.

THEOREM: *Suppose that the short-run per capita demand functions are locally linear in per capita expenditure,*

$$(14)\quad \bar{q}_{it} = b_{it} + f_i(P_t) + \gamma^i(P_t)\left[\bar{\mu}_t - f(P_t) - \sum p_{kt}b_{kt}\right]$$

*and* $b_{it}$ *is given by*

$$(15)\qquad b_{it} = \bar{b}_i^* + \beta_i\bar{q}_{it-1}$$

*Then the long-run per capita demand functions are given by*

$$(16)\quad h^i(P,\mu) = B^i(P) - \Gamma^i(P)\sum p_k B^k(P) + \Gamma^i(P)\bar{\mu}$$

*where*

$$(17)\qquad B^i(P) = \frac{\bar{b}_i^* + f_i(P)}{1-\beta_i}$$

$$(18)\qquad \Gamma^i(P) = \frac{\dfrac{\gamma^i(P)}{1-\beta_i}}{\sum \dfrac{p_k\gamma^k(P)}{1-\beta_k}}$$

With interdependent preferences one might expect the distribution of expenditure among individuals to play a significant part in determining the long-run con-

sumption pattern. Our theorem implies that this need not be true: under the specifications of interdependent preferences in which each individual is influenced by average past consumption, the long-run per capita demand functions depend only on per capita expenditure, and not on its distribution. The simplistic notion that interdependent preferences imply that the distribution of expenditure among individuals must influence the long-run consumption pattern is false. With some specifications it matters and with others it does not.

Now consider the specification in which each individual is influenced by the past consumption of the man above him in the hierarchy. Assuming only that the demand functions are locally linear in expenditure, these short-run per capita demand functions can be solved for the long-run equilibrium values, just as in the habit formation models. Thus, the long-run per capita demand functions, like their short-run counterparts, depend on per capita expenditure and on the range of the distribution of consumption. It is conceptually awkward to include the long-run consumption of individuals 1 and $R$ as arguments of the long-run per capita demand functions. We could use the habit formation results to determine the long-run consumption pattern of individual 1, and by substitution replace the $q^1$ by functions involving $\mu^1$, but since we cannot eliminate the $q^R$ in a similar manner, the form involving both $q^1$ and $q^R$ is more transparent.

Except for the terms involving the range, these demand functions are identical to those implied by the average past consumption specification, (16). In the limit, as the size of the community increases while the range of the distribution of consumption remains bounded, the terms involving the range approach zero and the demand functions approach those of the average past consumption specification. In this limiting case, the distribution of expenditure among individuals is not a determinant of the long-run consumption pattern. One must be careful, however, about applying this conclusion to large finite communities, since we would not expect the range to be independent of the size of the community.

Now consider the two-class model of interdependent preferences in which the tastes of both classes depend on the average past consumption of the upper class. Assuming linearity in expenditure and the same interdependence coefficients for both classes, the short-run per capita demand functions are easily determined. Since the behavior of the upper class is determined by habit formation, the long-run per capita demand functions of the upper class are given by the habit formation formula. Substituting this result into the short-run per capita demand function for the lower class yields the long-run consumption pattern of the lower class. It is easily seen that in the two-class model the long-run per capita consumption pattern does depend on the distribution of expenditure between classes.

In the two-class model the distribution of expenditure plays a significant role in determining the long-run consumption pattern. I shall not discuss "class" models with more than two classes, but one would expect distribution to matter if the number of classes is small. The two-class model is a polar case of the hierarchy specification corresponding to $R=2$. In the opposite polar case—when the number of individuals is very large and each individual's expenditure is small relative to the total—the per capita demand functions approach those of the average past consumption specification.

Perhaps because interdependent preferences were introduced in the context of the consumption function where they were generally supposed to lead to models in which the distribution of wealth or income influenced saving, it is widely believed that

interdependent preferences imply that the consumption pattern depends on the distribution of expenditure among individuals. After investigating a number of specifications of interdependent preferences, it is clear that whether or not the distribution of expenditure influences the long-run consumption pattern depends on the precise specification employed, and also, in some cases, on apparently extraneous factors such as the size of the community. Certainly specifications in which individual short-run demand functions depend on the distribution of past consumption need not imply short- or long-run per capita demand functions which depend on the distribution of expenditure among individuals.

## IV. Conclusion

I begin this section by tentatively evaluating the specifications of lagged interdependent preferences introduced above and comparing them with simultaneous interdependence, a specification in which preferences are influenced by other people's current consumption. I then discuss the dynamics of interdependent preferences and describe a specification which combines interdependence and habit formation. I conclude by describing a difficulty which interdependent preferences poses for welfare economics.

All of the specifications of linear interdependent preferences I have discussed have been introduced in conjunction with specific systems of demand functions. I have relied heavily on the linear expenditure system for exposition but the results apply to all systems of demand functions locally linear in expenditure. In most cases further extensions are possible, but they often involve substantial complications and I have not considered them in detail. The emphasis on simple and specific functional forms has been deliberate and reflects more than a desire to avoid mathematical difficulties. Empirical demand analysis requires specific functional forms, and I have emphasized classes of demand functions which are suitable for econometric work.

The various models of interdependent preferences begin with assumptions about the preferences or behavior of an individual, so they are suitable for interpreting budget study data. However, I have emphasized their implications for the per capita rather than the individual demand functions because the full ramifications of interdependent preferences depend on the reciprocal interaction of all individuals, and the outcome of this interaction manifests itself only in the long-run per capita demand functions. I have focused on demand functions locally linear in expenditure, so that the distribution of expenditure has no direct effect on per capita consumption patterns, and demand functions depend linearly on other people's consumption. The linearity assumption can be weakened, but the cost in additional complexity is likely to be high.

One of the most interesting substantive results to emerge from this investigation of interdependent preferences relates to the role of income or expenditure distribution as a determinant of long-run per capita consumption patterns. Although my initial intuition was that expenditure distribution should matter, there are a number of plausible cases in which it does not. Whether expenditure distribution matters depends on the detailed structure of the specification of interdependence employed; the notion that interdependence automatically implies a significant role for expenditure distribution is false.

Economists' views of interdependent preferences have been largely shaped by Duesenberry's simultaneous specification, and some will be uncomfortable with the lagged formulation proposed here. A significant advantage of lagged interdependence is that it implies a specific gradual adjustment path from one equilibrium to

another. Subject to the usual caveat about the length of the time period, I believe this is an important consideration. Unless the periods are very long, a specification which implies an immediate adjustment from one equilibrium to another must be supplemented by an *ad hoc* adjustment mechanism; *ceteris paribus*, a theory which implies a plausible adjustment process is preferable to one which does not.

Corresponding to each specification of pure lagged interdependence (i.e., lagged interdependence untainted by habit formation) is a related specification of simultaneous interdependence. If a specification of pure lagged interdependence is replaced by the corresponding model of simultaneous interdependence—that is, by the specification obtained by replacing each lagged value by the corresponding current value—then the demand functions of the simultaneous specification are identical with the long-run demand functions of the lagged specification. With simultaneous interdependence, there is no distinction between the short and long run, because the long run works itself out in a single period.

The relative merits of lagged versus simultaneous interdependence should be judged on their empirical usefulness, so it would seem that there is little to be said. My preference is for lagged interdependence because I believe that it is more plausible than simultaneous interdependence and because it is more tractable. One defect of simultaneous interdependence is its implication that the complete adjustment from one equilibrium to another takes place in a single period, but it is misleading to characterize simultaneous interdependence as a specification in which adjustment is instantaneous: the response to changes in prices and expenditure works itself out in a single period, but this is not a serious objection if the periods are long enough, say, a decade or a generation. On the other hand, the gradual adjustment implied by lagged interdependence seems more plausible for quarterly or annual data.

The dynamics of interdependent preferences are somewhat different from those of habit formation.[9] Even though the per capita demand functions appear identical, the dynamic analysis of the per capita demand functions should be based on the individual demand functions and these are not identical. In the linear expenditure system, regularity conditions require that each individual's expenditure be at least as great as the cost of his "necessary basket." With habit formation, this depends on his own lagged consumption, and the demand system is locally stable if the $\beta$ are all less than one. With interdependent preferences, if an individual is influenced by the average of everyone's past consumption, then regularity conditions will not be satisfied for individuals at the bottom of the expenditure distribution unless the $\beta$ are substantially less than one, but if an individual is influenced by the consumption of the man above him in the expenditure hierarchy, and the distribution of expenditure among individuals does not change drastically from one period to the next, then $\beta$ slightly less than one will imply stability.

Habit formation and interdependent preferences can operate together. In particular, suppose that each individual is equally concerned with everyone else's past consumption, so that $\beta_i^{rs} = \beta_i^r$, $r \neq s$, and that $\beta_i^{rr}$ is allowed to assume a distinct value. The implied individual demand functions are slightly messy, but the short-run per capita demand functions (derived under the assumption that all individuals have the same interdependence coefficients, the same marginal budget shares, and the same habit formation coefficients)

[9] The dynamics of the habit formation model with a one-period lag are analyzed in my 1970 article. M. D. McCarthy extends these results to a more general lag structure in which the memory coefficient need not be zero.

are of the same form as those obtained under the separate hypotheses of habit formation and interdependent preferences. It follows that the long-run per capita demand functions must also be of the same form as the long-run demand functions under either hypothesis alone. Hence, on the basis of the per capita demand functions, one cannot distinguish among habit formation, interdependent preferences, and a combination of the two: to do so, we must observe individual behavior.

Because we can distinguish between interdependent preferences and habit formation by observing individual behavior, these two hypotheses are not simply alternative names for the same device for introducing more flexibility into empirical demand equations. They are empirically distinct hypotheses whose implications happen to coincide for one particular type of data, but they are clearly distinguishable.

I conclude this section by discussing a problem which arises in using interdependent preferences as a basis for welfare judgments. If an individual is a price taker in markets for *all* of the goods which enter his preference ordering, then we can infer his preferences from his market behavior. In the present context, this condition will be satisfied only if he participates in markets in which he can pay other people to consume greater or smaller amounts of various goods. If the individual regards other people's consumption as given when he makes his market decisions, then we observe "conditional demand functions" of the form $q^r = h^r(P, \mu^r; q^{\bar{r}})$, where $q^r$ denotes the individual's intertemporal consumption vector and $q^{\bar{r}}$ denotes everyone else's intertemporal consumption pattern. From these demand functions one can infer the individual's preference ordering over $q^r$ conditional on the value of $q^{\bar{r}}$; we denote the implied conditional utility function by $V^r(q^r; q^{\bar{r}})$. The complete utility function, $U^r(q)$, must be consistent with the conditional utility function; this implies

$$U^r(q) = W[V^r(q^r; q^{\bar{r}}), q^{\bar{r}}]$$

where $W_1 > 0$, but the individual's market behavior implies no further restrictions on $W$.

This is disturbing because it implies that with interdependent preferences we cannot base judgments about individual welfare on the preference ordering revealed by market behavior; to make judgments about the welfare of an individual we need more information. To forestall confusion, it should be emphasized that this conclusion is unrelated to problems of cardinal utility or interpersonal comparisons of utility. Even if we are concerned only with judgments about the welfare of individual $r$ on the basis of his own ordering of social consumption vectors, we cannot identify this ordering without more information than market behavior provides.

## REFERENCES


J. S. **Duesenberry**, *Income, Saving, and the Theory of Consumer Behavior*, Cambridge, Mass. 1949.

**W. Gaertner**, "A Dynamic Model of Interdependent Consumer Behavior," mimeo. 1973.

**W. M. Gorman**, "On a Class of Preference Fields," *Metroeconomica*, Aug. 1961, *13*, 53–56.

**M. D. McCarthy**, "On the Stability of Dynamic Demand Functions," *Int. Econ. Rev.*, Feb. 1974, *15*, 256–59.

**R. A. Pollak**, "Habit Formation and Dynamic Demand Functions," *J. Polit. Econ.*, July/Aug. 1970, *78*, 745–63.

———, "Habit Formation and Long Run Utility Functions," *J. Econ. Theory*, forthcoming.

**S. J. Prais and H. S. Houthakker**, *The Analysis of Family Budgets*, Cambridge 1955.

# Years and Intensity of Schooling Investment

By Arleen Leibowitz*

In estimating human capital earnings functions it has been assumed in the literature that during schooling years 100 percent of gross potential earnings are invested in acquiring human capital, while in on-the-job training this percentage is smaller and declines with age. The assumption that the proportion of gross potential earnings invested ($K$) is unity during schooling years is useful in empirical work. It allows the identification of an estimate of the rate of return to schooling from a regression of earnings on years of schooling.

This paper argues that contrary to the usual assumptions in the literature, the percentage of gross earnings invested may fall below 100 percent well before schooling is ended, that this percentage is likely to be correlated with years of schooling, and thus the assumption that $K=1$ yields only a biased estimate of the rate of return to schooling. In this paper, a unique data set, the Terman sample, is used to calculate earnings functions. Because the Terman sample contains data on the costs of schooling (including tuition and net opportunity costs) as well as earnings over a 20-year period for a sample of 822 men, two of the limitations of census-type data can be overcome. First, returns to schooling costs rather than schooling years have been estimated. The percentage of gross potential earnings invested in schooling has been calculated and related to earnings. In addition, problems of estimating lifetime income profiles from cross-section data are overcome with data on lifetime earnings of the Terman sample. Lifetime earnings of men in the Terman sample, who were all in the top 1 percent of the IQ distribution in childhood, are compared with earnings of average ability males, and it is argued that much of the return to high ability results from more intensive schooling and early entry into the labor force.

Section I provides a review of the literature and argues that it is necessary to know the intensity of schooling investment ($K$, the proportion of gross potential earnings invested) in order to calculate rates of return to human capital investment. Section II utilizes the Terman sample to calculate intensities of investment and to estimate earnings functions. In Section III earnings of men in the Terman sample are compared with earnings of men of average ability.

## I. Intensity of Investment

### A. *The Bias in Earnings Functions not Specifying the Intensity of Investment*

The rate of return is a central parameter in human capital analysis. Although in the theoretical development of human capital,

* Brown University and National Bureau of Economic Research. The research assistance of David L. Lindhauer and Arun K. Mukhopadhyay is gratefully acknowledged. I also benefited from the helpful comments of G. Becker, B. Chiswick, B. Friedman, J. Mincer, S. Polachek, P. Wachtel, and the managing editor of this *Review*. This research was sponsored by a grant from The National Institute of Child Health and Human Development to the National Bureau of Economic Research and later by Grant HS01623-01 from the National Center for Health Services Research. The following paper is not an official National Bureau publication since the findings reported herein have not yet undergone the full critical review accorded the National Bureau's publications, including approval of the Board of Directors.

the rate of return on an investment is defined as that discount rate which equates the present value of returns over the lifetime of a given individual with the present value of costs of that investment (see, for example, Gary Becker, 1964, p. 39), the empirical work has generally had to rely on cross-section data. Since lifetime income histories for given individuals have only recently become available, cross-section data on people of various ages have been used to approximate lifetime earnings of an individual (see, for example, Becker, 1964, and Giora Hanoch).

A less laborious method of estimating rates of return from cross-section data was proposed by Becker and Barry Chiswick. This method also made possible simple controls for factors such as ability and other human capital investments, which were difficult to incorporate in the discounted present value approach. Becker and Chiswick reformulated the analysis in a way that allowed the rate of return to be estimated as a parameter in a regression where earnings was the dependent variable and the investment period was the independent variable. They define $K_{ij}$ to be the percent of potential earnings that individual $i$ invests in year $j$ $(0 \leq K_{ij} \leq 1)$. Then, abstracting from depreciation, gross earnings in year t can be expressed as a function of the initial stock of human capital $E_{0i}$, the series of $K_{ij}$, and the rate of return on these investments:

$$(1) \qquad E_{it} = E_{0i} \prod_{j=1}^{S} (1 + K_{ij} r_{ij})$$

But since $K_{ij} \leq 1$ and $r_{ij}$ is small, this can be written:

$$(2) \qquad \ln E_{it} = \ln E_{0i} + \sum_{j=1}^{S} K_{ij} r_{ij}$$

Assume either that all available time is invested in schooling during the "schooling years," and there are no direct costs of schooling such as tuition so that the costs are the wages foregone or, alternatively, that part-time earnings exactly offset direct schooling costs. Then $K_{ij} = 1$ for all $i$ and $j$. Assuming also that $r$ is constant over individuals and years of schooling, and that there is no postschool investment, then we have

$$(3) \qquad \ln E_{it} = \ln E_{0i} + rS_i$$

where $S_i$ is the number of investment years where $K_{ij} = 1$.

This formulation appears to allow the estimation of the rate of return to schooling investments from the relationship of a simple function of earnings and years of schooling. A further advantage of this procedure is that it allows one to hold constant in the estimating equation other forms of capital accumulation, such as ability or postschool investment. The literature which has grown up in these two areas has indeed relied heavily on the semilog functional form derived by Becker and Chiswick.[1]

The period of specialization in human capital investment, when all resources are devoted to acquiring human capital and $K = 1$, has generally been assumed coincident with the period an individual is enrolled in school (see, for example, Yoram Ben-Porath, Thomas Johnson). An exception is the work of William Haley, who explicitly allows $K$ to fall below unity before formal schooling is terminated.

Because of the prevalence of part-time schooling and market work undertaken during the schooling years, $K$ need not equal unity. However, simply dividing the coefficient $r$ by $K$ will not yield an unbiased estimate of the rate of return. A proof of this statement follows.

It is easy to show that even if $K_{it} = 1$ on average, unless $K_i$ is independent of $S_i$, $r$ in equation (3) is a biased estimate of the

[1] See, for example, Zvi Griliches and William Mason, John Hause, and Jacob Mincer (1974).

rate of return to schooling.[2] Let $\overline{K}$ be the average percentage of gross earnings invested in schooling over all individuals and all years of schooling, and $h_{it}$ be the deviation from this average for the $i$th individual in the t-th investment year.

$$K_{it} = \overline{K} + h_{it}$$

$$\sum_{t=0}^{S} K_{it} = S_i\overline{K} + \sum_{t=1}^{S} h_{it}, \qquad \sum_{it} h_{it} = 0$$

Let
$$h_i = \frac{\sum_{t=0}^{S} h_{it}}{S_i}$$

be the $i$th person's average annual deviation in percent of gross earnings invested per school year. Then equation (3) can be rewritten:

$$log\ E_{it} = log\ E_{0i} + r_1 S_i \overline{K} + r_2 S_i h_i$$

where $r_1$ is the rate of return on years of average intensity schooling and $r_2$ is the rate of return on deviations from the average intensity. From this equation, it is clear that if $S_i h_i$ and $S_i$ are correlated, one cannot in general identify the rate of return to schooling investments from a regression like equation (3) of *log* earnings on years of schooling alone, even if the average intensity of investment $\overline{K}$ is known. The omission of intensity of schooling in previous empirical estimates of earnings functions has led to biased estimates of the rate of return to schooling, for two reasons. First, the fact that average intensity $\overline{K}$ may differ from unity is ignored. The estimated coefficient $r$ is biased upward (downward) as $\overline{K}$ is greater than (less than) unity. Second, the correlation between $S_i h_i$ and $S_i$ is ignored leading to a bias in the estimated coefficient of $r_2 b$, where $b$ is the regression coefficient of $S_i h_i$ on $S_i$. Ignoring the positive relation between intensity and years of schooling imparts an upward bias to the rate of return estimated from equation (3). In Section II we estimate the correctly specified equation.[3]

### B. *The Interpretation of Years of Schooling in Earnings Functions*

Because data on the intensity of schooling investment have not been readily available, earnings functions of the form of equation (3) have been estimated. The relationship between earnings and years of schooling has been amply demonstrated. What interpretation can this relation be given? It is neither the market rental on human capital, which plays an important role in much of the theoretical analysis (see Ben-Porath), nor is it a rate of return on investment costs since years of schooling accurately reflect neither human capital stocks nor investment costs. To clarify these issues, consider two alternative ways of decomposing $E_{it}$ earnings of individual $i$ in year t. (Hereafter the subscript $i$ will be dropped.) The approaches are the market rental and rate-of-return calculations. In the market rental analysis earnings are the scalar product of a vector $H$, of quantities of various kinds of human capital possessed by the individual at time t, and the vector of markets rents, $W$, available on these various kinds of capital.

$$E_t = \sum_{j=1}^{M} W_{tj} H_j \tag{4}$$

A perfectly competitive market for human capital services is assumed and $W_{tj}$ does not vary across individuals.

An alternative approach is to consider the relationship between earnings and the cost of producing the stock of capital

[2] Chiswick, pp. 2–13a, notes that some of the variance in income may be due to the covariance of $rK$ and $S$.

[3] The derivation of equation (3) by Mincer (1974, pp. 9–11) is not dependent upon the assumption of the independence of $h_i$ and $S$. However, the Mincer proof requires that all individuals have identical present values of earnings.

which is rented in the market. Assume a production process whereby a vector of inputs $I$ is transformed into a vector of human capital outputs $H$. The production process may be characterized by the matrix $A$ which defines the linear transformation of input vector $I$ into the vector of human capital outputs $H$. The cost per unit of input is given in the vector $C$.

$$H_j = \sum_{k=1}^{N} A_{jk} I_k \tag{5}$$

$$E_t = \sum_{k=1}^{N} r_{tk} C_k I_k \tag{6}$$

where $r_{tk}$ is the rate of return for each kind of capital in year t. Assume constant returns to scale for all types of capital investments so that marginal returns equal average returns and assume also that returns are constant over time. Since an optimizing individual would equate returns on various types of capital at the margin, we can write earnings as the product of a constant rate of return on the various investments:

$$E_t = r_t \sum_{k=1}^{N} C_k I_k \tag{7}$$

Thus $r_t$ can be defined as the individual's rate of return in year t for his human capital investments:

$$r_t = \frac{\sum_{j=1}^{M} W_{tj} H_j}{\sum_{k=1}^{N} C_k I_k} \tag{8}$$

The problem in estimating equation (3) above is that years of schooling correspond most closely to a quantity of inputs $I$, whereas the rate of return calculation requires the denominator of equation (8) of which $I$ is only one part. Years of schooling is equally inappropriate for estimating the market rental of human capital $W$ (as in Johnson, p. 551) since the transformation of inputs into human capital outputs described by (5) is ignored.

### C. *The Relationship of Intensity to Years of Schooling*

While it is easy to show (see Ben-Porath, 1970) that the proportion of gross potential earnings that an individual allocates to investment decreases over his lifetime, the relationship between years and intensity of schooling across individuals is not obvious. If individuals faced identical marginal productivity and supply of funds schedules, and they chose schooling attainment randomly, we would then expect a negative correlation between intensiveness (resources per year) and extensiveness (the number of years) of schooling investment, since intensiveness and extensiveness would be substitutes in producing the (identical) total stock of human capital. However, individuals are not identical, and it seems likely that some of the factors which affect the choice of years of schooling affect the desired intensity of schooling in the same direction, leading to a positive correlation between the two.[4]

Holding ability and other factors affecting the marginal productivity of schooling constant, consider the supply curve of funds which relates the cost of financing the investment to the intensity of the investment process in a given year of schooling. For any given number of years of schooling, more intensive investment can be financed only at greater cost per unit of investment. In Figure 1 the supply of funds curve $K_2$ referring to greater intensity of investment is above curve $K_1$,

[4] Several recent papers have taken the length of the period of specialization as endogenous (Haley, Lee Lillard, Dudley Wallace and L. A. Ihnen), but most except Haley have taken the specialization period as coincident with years of schooling. Haley considers the case where the only input to investment is time, and he shows that quantity and the proportion ($K$) of human capital stocks allocated to producing more human capital declines strictly monotonically, after the period of specialization when $K=1$ (p. 934).

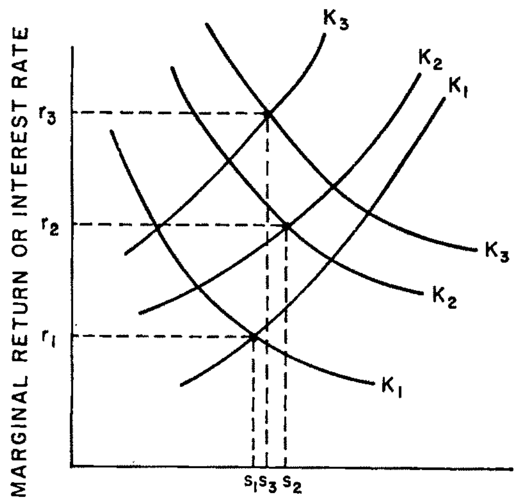



FIGURE 1. YEARS OF SCHOOLING

indicating that for any given number of years of schooling, higher interest rates must be paid to finance a greater proportion of gross potential earnings. As Becker has pointed out the funds "available to any person from the cheaper sources are usually rationed since the total demand for funds tends to exceed their supply" (1967, p. 9). Thus as increasing amounts of funds are required within a given year, students shift from gifts from relatives and grants from government or schools (which usually are subject to a maximum amount per school year) to subsidized loans from government or universities (also subject to a maximum amount per year). Finally they may turn to loans at commercial interest rates or to reduced consumption.

In Figure 1 we also see that the interest rate required to finance schooling will be a positive function of the total number of years of schooling at any given intensity. This is so because the dollar outlay corresponding to any given value of $K$ rises with schooling, which increases earnings. Secondly, since the total amount invested is rising with $S$, the interest rate should too.

Next consider the relationship between the return and the intensity of investment of a marginal year of schooling. This will determine the demand for schooling intensity for one additional year. Assume that the amount of human capital produced in a school year is a positive function of the intensity of schooling, and that employers are able to distinguish the greater human capital stocks of persons who invest more intensively.[5] Psychologists have shown that even during the time that students are ostensibly learning, they may be diverting a considerable proportion of their resources to noninvestment uses (see, for example, Lorin Anderson). These studies using both experimental and naturalistic classroom situations find that the percentage of "time on task" ($K$) is positively related to the amount learned (human capital developed). Although additional resources invested imply additions to the stock of human capital, because of the strict limit on the amount of time available to an individual, diminishing returns to producing additional capital are assumed. Figure 1 indicates that those with intensity of investments $K_1$ would get $S_1$ years of schooling and a rate of return (or borrowing cost) of $r_1$, while schooling intensity $K_2$ would be associated with $S_2$ and $r_2$. Now it is clear from Figure 1 that, holding ability constant, either positive or negative correlation between $S$ and $K$ is possible.

Ability affects the productivity of time in schooling positively, so that for any given intensity, the more able students' demand for schooling curve lies above that of the less able student. On the supply side the more able students are more likely to obtain grants or low-cost loans from their universities or from the government, and

[5] If this were not true, students could maximize the rate of return on the $j$th year of schooling by investing the minimum amount which would allow them to claim they had "completed" $j$ years of school.

to come from higher income families, who can more easily finance schooling, so that their supply of funds curves lie to the right of less able students. Both supply and demand forces thus lead the abler student to demand more years of schooling. Confronted by schooling "careers" which meet the demand of the average student, and call for no more than five or six years of training beyond the B.A., abler students can adapt by investing more intensively. Thus the rate of return on the years of schooling of abler students tends to overestimate the rate of return on the cost of schooling, since they have an incentive to invest more intensively.

Holding ability constant, there is no clear prediction about the sign of the correlation between years and intensity of schooling. Thus the direction of bias in rate of return estimates based on the assumptions $K=1$ and independent of $S$ is an empirical question. Bodies of data with cost of schooling and information on earnings are rare, which is one reason for the widespread use of equation (3) as an estimating function. However, the Terman data, which are described in the following section, can provide an illustration of some of the points raised above.

## II. Earnings Functions Based on Cost of Investment

### A. *The Terman Sample*

The Terman sample consists of persons whose measured IQ fell within the top 1 percent of the IQ distribution when they were tested as children in 1921–22.[6] This sample of 1,528 persons with IQs of 135 and above was originally selected from the population of California school children by Lewis Terman, a Stanford University psychologist. Data on the earnings and schooling of these high ability individuals were updated by resurveys in 1929, 1940, 1950, and 1960.

In the 1940 questionnaire the Terman sample members were asked to supply the name of the college they attended, the amount of scholarships and assistantships they had received and their total earnings as undergraduates. We developed a tuition series for six different colleges named for the years 1921 to 1940. Of the individuals who reported cost data, 95 percent attended the six colleges for which tuition data were obtained. Cost data were supplied by 85 percent of the sample members who attended college. Using data on age at high school graduation and age at receipt of B.A., we developed an estimate for each individual of the direct costs of his undergraduate schooling. Net indirect costs were estimated by using the data on scholarships, assistantships, and earnings. Opportunity costs were estimated to be \$1,092 per year in 1927 dollars. This figure was arrived at by using data from Terman (III, pp. 137–38) which showed that average weekly compensation for full-time jobs in 1927 was \$21 for male college students in the Terman sample and \$22 for men in the sample who were not in school. However, if actual earnings exceeded this estimate of opportunity costs, actual earnings were used as the measure of full-time earnings capacity.[7] All figures were put in constant (1947) dollars, the total cost of schooling was calculated as:

$$
\begin{aligned}
\text{Total Cost} = {} & \text{Tuition} + \text{Opportunity Cost} \\
& - \text{Total Scholarships} \\
& - \text{Total Assistantships} \\
& - \text{Earnings}
\end{aligned}
$$

Then $$K = \frac{\text{Total Cost}}{\text{Opportunity Cost}}$$

Table 1 presents mean values for the

[6] I am indebted to Robert Sears and Melieta Oden at Stanford University for making these data available.

[7] This accounts for the opportunity costs of college dropouts exceeding opportunity costs of graduates.

TABLE 1—AVERAGE COMPONENTS OF COLLEGE COSTS BY FINAL SCHOOLING LEVEL (in 1947 dollars)

| Schooling Level in 1940: | $12<S<16$ | $S=16$ | $S=16$ | $S>16$ |
|---|---|---|---|---|
| Schooling Level in 1950: | | $S=16$ | $S>16$ | |
| Average Yearly Schooling Costs | | | | |
| Tuition | 176.8 | 223.2 | 285.1 | 247.0 |
| Scholarships | 100.4 | 20.1 | 12.9 | n.a. |
| Assistantships | 202.6 | 1.08 | 20.0 | n.a. |
| Earnings | 702.2 | 450.4 | 239.0 | 330.9 |
| Opportunity Cost | 1648 | 1485 | 1472 | 1476 |
| $K$ = Percent of Opportunity Cost Invested | .55 | .84 | 1.01 | n.a. |
| Number of Observations | 87 | 134 | 19 | 248 |

*Source:* Calculated from the Terman sample.

entire sample of the cost components of undergraduate schooling as well as for $K$. For the entire sample with complete cost data, $K=.75$. Table 1 indicates that the percent of potential earnings per year invested in schooling increased with final schooling level attained.[8] For persons with less than a B.A. in 1940, the annual percent of gross earnings potential invested in schooling, $K=.55$. For those with a B.A. in 1940 who had not increased their schooling level by 1950, $K=.84$, while those who earned advanced degrees after 1940 invested 101 percent of gross annual earnings in their undergraduate schooling.[9] Thus, only the group who went on to graduate studies spent 100 percent of gross potential earnings in investment during their undergraduate years. Unfortunately, because graduate and undergraduate scholarships and assistantships awarded up to 1940 were not segregated in the data, it is not possible to calculate $K$ for persons with graduate training before 1940. Table 1 shows that those who completed a higher level of schooling spent more money per year for tuition and diverted less of their available time to the labor market during their schooling years.[10]

### B. *Earnings as a Function of Years and Cost of Schooling*

For the subsample of individuals on whom cost data are available, we first estimate the relationship between earnings and years of schooling in order to have a baseline with which to compare earnings functions based on costs.[11] The methodology is to combine income data from the three dates (1940, 1950, 1960) for all males on whom we had cost and income data to estimate a lifetime earning function. For this group each additional year of college increases wages by 6.9 percent, while each year of graduate schooling (obtained after 1940) increases income by 4.7 percent. (See Table 2, col. 1.)

If we were to assume that years and intensity of schooling were uncorrelated, we could derive an estimate of the rate of

[8] The anomalous result that those who did not complete their schooling until after 1940 had greater annual investments in undergraduate training than those who already had some graduate work by 1940 may be due to a cohort effect, since the former group is likely to be younger and have started to school in slightly better times.

[9] These estimates are similar to those computed by Paul Wachtel.

[10] The lower earnings of those who completed more years of schooling imply less time in the labor market. The earnings data imply that if they spent the same amount of time working, their wages were only 34 to 64 percent of those earned by students who completed less schooling. This seems unlikely even though the students who did not complete college may have failed to do so because of higher opportunity costs.

[11] A subsample of 306 individuals who had no more than 16 years of schooling in 1940 was used. Some had obtained graduate schooling by 1950 and 1960.

TABLE 2—ANNUAL EARNINGS AS A FUNCTION OF HUMAN CAPITAL INVESTMENTS

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Intercept | 3.39 | 3.41 | 3.36 | 3.40 |
| | (30.7) | (31.6) | (30.7) | (23.2) |
| College Years | .069 | | .034 | .030 |
| | (5.30) | | (1.91) | (1.03) |
| Graduate Years | .047 | .094 | .076 | .086 |
| | (1.59) | (3.25) | (2.17) | (2.40) |
| Total Cost[a] | | .065 | | |
| | | (5.60) | | |
| Experience | .049 | .051 | .051 | .048 |
| | (4.29) | (4.47) | (4.55) | (3.55) |
| Experience$^2$ | $-.0005$ | $-.0005$ | $-.0005$ | $-.0004$ |
| | $(-1.80)$ | $(-1.98)$ | $(-1.99)$ | $(-1.28)$ |
| $S \times K_{Tuition}$[b] | | | .231 | .235 |
| | | | (3.81) | (3.80) |
| $S \times K_{Scholarship}$ | | | 3.41 | 3.92 |
| | | | (.20) | (.23) |
| $S \times K_{Assistantship}$ | | | $-19.1$ | $-21.5$ |
| | | | $(-.88)$ | $(-.99)$ |
| $S \times K_{Earnings}$ | | | $-.375$ | $-.247$ |
| | | | $(-.11)$ | $(-.07)$ |
| $R^2$ | .148 | .151 | .164 | .172 |
| $N$ | 901 | 901 | 901 | 645 |

*Note:* Dependent Variables is *Log* of Earnings in 1947 Dollars

[a] Total Cost $= \sum_{j=1}^{S} K_{ij}$

[b] $K_{Tuition}$ is the percentage of gross potential earnings spent on tuition; $K_{scholarship}$, $K_{assistantship}$, and $K_{earnings}$ are the percentages of gross potential earnings covered by scholarships, assistantships, and earnings.

return to schooling by dividing the schooling coefficient in equation (1) by .75, the sample average value of $K$.[12] That estimate of the rate of return is 9.2 percent. However, since we know years of schooling and $K$ are positively correlated, this is an overestimate, as shown above.

The availability of cost data allows a direct estimate of $r$. In column 2 of Table 2, the results of regressing earnings on sum of the elements of $K$ are presented. The estimated rate of return of 6.5 percent is surprisingly close to the estimate derived from equation (1) with the assumption that $K=1$. This occurs because of two offsetting biases. The coefficient on years of schooling in equation (1) is biased downward because no account is taken of the fact that $K<1$. However, ignoring the positive correlation between $K$ and $S$ imparts an upward bias to the coefficient. Note also that the coefficient on years of graduate schooling rises in magnitude and significance, since those who invested more in their undergraduate schooling were more likely to obtain graduate schooling. Without accounting for the greater $K$ of these people, the coefficient on college years was biased upward, and consequently, the coefficient on years of graduate schooling was biased downward.

Since a B.A. is a prerequisite to most graduate programs, this suggests that more intensive undergraduate work may provide its return by allowing entrance to a gradu-

[12] This reasoning has been applied, for example, by Lewis Solomon, pp. 28–29. Chiswick, pp. 3–15, applies a similar procedure.

ate program. The coefficients in equation (2) should not be taken to mean that the rate of return to graduate school is twice as great as the rate of return to undergraduate schooling, for reasoning as we did above, the coefficient on years of graduate school cannot be interpreted as a rate of return. And if the intensity and years of graduate schooling are positively correlated, the coefficient of equation (2) is an overestimate of the rate of return.

The relationship between earnings and the various components of average $K$—tuition, fellowships, foregone earnings, etc., can be determined. $K_{ij}$ can be decomposed into the fraction of schooling investment accounted for by opportunity costs and the fraction accounted for by the other components of $K$, as is shown below.[13] Since:

$$r\sum_{j=1}^{S_i} K_{ij} = rS_i + r\left[\frac{T_i - F_i - A_i - E_i}{O_i/S_i}\right]$$

$T_i$ = average yearly tuition for $i$th individual

$F_i$ = average yearly scholarship for $i$th individual

$A_i$ = average yearly assistantship for $i$th individual

$E_i$ = average yearly earnings for $i$th individual

$O_i$ = average yearly opportunity cost

$S_i$ = years of college

The results of estimating this equation are shown in column 3 of Table 2.

The hypotheses about the signs of these variables depend on their effects on the amount of human capital acquired per school year. Because it is assumed that greater yearly tuition will correspond to greater school inputs per academic year, the partial effect of tuition is expected to be positive.[14] The percent of opportunity cost covered by earnings is a proxy for the amount of time allocated to nonacademic uses per school year. Thus undergraduate earnings should be negatively related to human capital acquired, and thus to future income. However, undergraduate earnings may be a measure of market ability, and therefore positively related to future earnings. Assistantships usually require working in the university, therefore allocating time away from studies, and are expected to be negatively related to future earnings. However, since the most able students are the recipients of assistantships, the coefficient in the regression may be biased toward zero. Scholarship holders, like recipients of assistantships, should be an exceptionally able group. However, since scholarships do not usually require taking a university job, their net effect on later earnings should be positive.

In column 3 of Table 2 we see that although all the components of schooling intensity have the expected signs, only years and tuition intensity have significant effects on later earnings. Although scholarship holders had higher earnings than assistantship holders, or those with jobs, these differences are not significant. We can infer, at least for the very able, that a little hard work never hurt anybody (in terms of future earnings).

When a comparable equation is estimated for the 645 observations on students with at least some college, the results are quite similar in terms of the proportion of gross earnings invested in tuition, assistantships, scholarships, and earnings (see col. 4). However, the opportunity cost component of investment (represented by college years) drops in significance,

[13] A similar approach has been taken by Wachtel, p. 5a.

[14] If tuition is related positively to parental wealth and not at all to the quantity of school inputs purchased, its expected sign would still be positive. In this context, the two hypotheses cannot be differentiated.

TABLE 3—RATIO OF MEAN INCOMES OF TERMAN SAMPLE TO CENSUS BY AGE AND EDUCATION: 1940, 1950, 1960

| | High School 4 Years | | | College 1–3 Years | | | College 4 Years | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1940 | 1950 | 1960 | 1940 | 1950 | 1960 | 1940 | 1950 | 1960 |
| **I. Age** | | | | | | | | | |
| 22–24 | 3.7 | | | 3.6 | | | 3.8 | | |
| 25–29 | 3.0 | | | 2.8 | | | 2.3 | 1.5 | |
| 30–34 | 2.3 | 1.8 | | 1.9 | 2.7 | | 1.9 | 1.6 | |
| 35–44 | 2.2 | 2.7 | 1.1 | 1.4 | 2.0 | 1.8 | 1.6 | 1.7 | 1.4 |
| 45–54 | | 1.7 | 1.9 | | 1.2 | 1.6 | | 1.6 | 1.2 |
| 55–64 | | | 1.1 | | | 1.1 | | | 1.2 |
| **II. Experience:[a] Adjusted-Age** | | | | | | | | | |
| 23–27 | 3.3 | | | 2.7 | | | 2.4 | | |
| 28–32 | 2.4 | | | 2.2 | 1.2 | | 1.8 | 1.3 | |
| 33–42 | 2.0 | 2.0 | | 1.9 | 2.1 | 1.3 | 1.4 | 1.7 | 1.2 |
| 43–52 | | 3.5 | 1.7 | | 1.6 | 1.9 | | 1.6 | 1.3 |
| 53–62 | | | 2.4 | | | 1.5 | | | 1.2 |

*Source:* Appendix.

[a] The ratio of earnings of Terman sample members of the age indicated are given relative to earnings of men who are two years older, but have the same number of years of experience in the labor market.

while the coefficient of graduate years rises.

### III. Ability, Intensity, and Earnings

Ability is a major factor shifting the marginal productivity and marginal cost curves for intensity and years of schooling. To see how high ability affects the market rental on equivalent years of schooling, earnings of the high ability males from the Terman sample are compared with earnings of males reported in the U.S. Census in this section.

Table 3 reports ratios of mean income earned by Terman sample members to mean incomes of males reporting income in the 1940, 1950, and 1960 census. The data for the Terman sample were collected from questionnaires administered in 1940, 1950, and 1960 and the data for the three years are defined as follows:

1940: Monthly wage and salary income, times 12, for sample members who were employed full time;
1950: Annual employment income in 1949 for sample members employed full time;
1960: Annual employment income in 1959 for sample members employed full time.

The income data are presented in the Appendix, along with census data for comparison. Ratios of the incomes of Terman subjects to incomes of census respondents are presented in Table 3.

In all census years for all education and age groups, average incomes of the Terman sample exceeded the mean incomes of the population at large. Where comparisons are possible of similar age-education groups in 1940, 1950, and 1960, no marked trends are discernible. That is, the ratio of Terman incomes to census incomes for individuals of a given age and education does not rise or fall in a systematic way between 1940 and 1960.

However, if a given cohort is followed through the three census years, the percentage by which their income exceeds census income shows a pronounced downward trend. (There are decreases in 16 of 18 cases.) The null hypothesis that there is no trend in the ratios (i.e., the probability of decrease $=.5$) can be rejected at the .001 level of significance, using a binomial test. Unusually high ratios in 1940 may be due to a sampling bias in the 1940 census, since earnings were reported only

for persons with no more than \$50 of other income. Because of this restriction many persons with high earnings may have been excluded, leading to a downward bias in the earnings data reported by the Census. However, since the comparisons in 1940 cover primarily the early years of the life cycle, the census sample may not be overly biased. Earnings of the Terman sample may also have been overestimated due to the assumption that sample members worked a full year. However, the trend in income ratios over time is not solely an artifact created by data deficiencies, since the ratios of Terman to census incomes also falls between 1950 and 1960. Furthermore, there is a decrease in the ratio with age within education groups in a given year in 18 of 22 cases. (The probability of obtaining this result when $p=.5$ is .002.)

Figure 2 illustrates the situation. While the earnings of Terman sample members are always greater than mean earnings from the Census, they rise at a less rapid rate.

The finding that both over time and within a given year, relative earnings of high ability males fall with age is surprising and contrasts with the findings for relative earnings over the life cycle for persons who differ in the amounts they

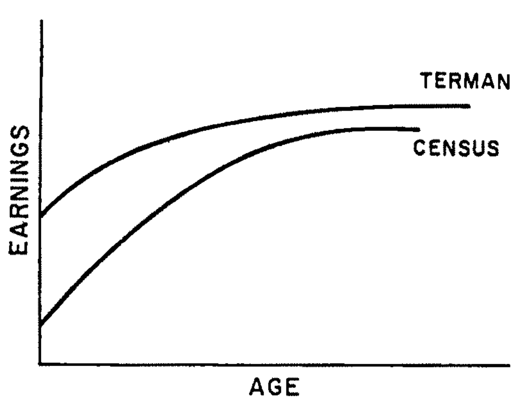



FIGURE 2. LIFETIME EARNINGS OF TERMAN AND CENSUS SAMPLE OF SAME SCHOOL AND AGE COHORT

possess of another kind of human capital—years of schooling. Mincer finds (1974, p. 70) that relative wages of the more schooled increase with age, although experience-income profiles are nearly parallel for *log* of weekly earnings and tend to converge for *log* of annual earnings. Using equation (2) to compare earnings of groups $m$ and $n$ with different initial endowments of ability, $E_0$, but the same number of years of schooling, we can write:

$$(9)\quad ln\left(\frac{E_{mj}}{E_{nj}}\right) = (ln\ E_{0m} - ln\ E_{0n}) + r_s \sum_{t=0}^{S} (K_{tm} - K_{tn}) + r_t \left[ \sum_{t=S+1}^{j-1} (K_{tm} - \delta_{tm}) - (K_{tn} - \delta_{tn}) \right] + ln\,(1 - K_{jm}) - ln\,(1 - K_{jn})$$

Schooling investments, which occur until time $S$, have been separated from postschool investments. Net postschool investments (over depreciation) earn a return of $r_t$, while the return on schooling investments is $r_s$.

Since the difference between the initial endowments and the schooling investments of the two groups remains constant over time, the only source of variation in their relative incomes over the life cycle is in the term $(K_t-\delta_t)$, reflecting postschool investments and depreciation.

In this model postschool investments are financed by lower net (observed) earnings. Thus the less steeply rising earnings profile of the Terman group implies that a vis-à-vis the census group either their postschool investments are a smaller proportion of gross potential earnings (although the value of these investments may be greater than those of the census group) or that the depreciation on their human capital is greater.

Mincer concludes (1974, ch. 2) that persons with more human capital in the form

of years of schooling also have greater money expenditures for postschool investments. But he finds a zero or negative correlation between time spent in schooling and in postschool investments. The census and Terman earnings profiles converge even more rapidly than those of various schooling groups. There are three possible explanations of this convergence:

1) Percentage of gross earnings invested in postschool investment is lower for the Terman group. The Terman group was selected because of high IQ scores; and the one thing IQ scores predict best is ability to succeed (i.e., acquire human capital) in school. Thus the "ability" criterion may be defined in such a way as to have a nonneutral effect on productivity in producing human capital in school and on the job. High IQ persons may invest more intensively during school years because they are relatively efficient at acquiring capital through schooling, while others invest more intensively on the job.

2) A more rapid rate of depreciation among the very able is another possible source of their slower rate of increase in net earnings. The impact of family characteristics which are known to be associated with higher earnings do affect initial earnings to a greater extent than later earnings. In the Terman sample (see my 1974 paper) the positive impact of family income on the subjects' earnings decreases over time. Thus the shrinking differential between high and average ability persons may be due to the falling value of family contacts with which the Terman sample was more heavily endowed than others in their age cohort.

3) Although men with the same numbers of years of schooling and continuous labor force experience should have the same number of years of labor force experience, this may not be the case in this sample. Because Terman sample members showed high IQs while still in grade school, many were accelerated in their schooling. Nearly half (49.3 percent) of the sample graduated from high school before age 17 and the mean age at college graduation was 21.5,[15] whereas the mean age for all college graduates was calculated as 24 by Hanoch.

Because of the greater intensity of their early schooling, Terman sample members have approximately two more years of market experience for any given schooling level. This would mean that the earnings function of Terman subjects diagrammed in Figure 2 should be shifted to the left to compare with persons of equivalent experience from the Census. When incomes are again computed, we see that the ratios decline less steeply over time and with age. (See Panel II, Table 3). Using "experience adjusted" comparisons, the proportion of declines from one census year to the next remains significantly different from .5 at the 1 percent level, but the fraction of declines within age groups is only significantly different from one-half at the 22 percent level. Before correcting for their accelerated schooling, the high ability sample earned from 10 percent to 280 percent more than persons with average ability and the same age and schooling. Much of the difference in lifetime earning patterns disappears when the data are adjusted to account for the greater intensity of early schooling of the high ability sample.

The much greater earnings of high ability males shown in this sample for given schooling levels contrasts with rather modest effect of ability on earnings found by Griliches, Hause, and Paul Taubman. Given the very high IQ level in the Ter-

[15] Terman, Vol. III, pp. 264–69. The modal and median value was 21.5 (my calculation) for college graduation. The mean age at high school graduation was 17.0 years (Terman, Vol. III, p. 266). Within the Terman sample the brighter students were accelerated. The students with IQs above 170 finished high school at 16.2 years on average. Of this group the 39 who completed college did so at an average age of 20.9 years.

man sample, this may imply a strong nonlinear effect of ability on earnings. Some of the differential between earnings of the high ability sample and those of average ability may be due to correlation of ability with other factors such as higher quality of schooling or higher family incomes.[16] It is not possible to test these alternatives given the lack of a detailed longitudinal data for persons of average ability comparable to the Terman study.

However, one of the causes of differences in the rates of change of income does appear to be related to different ages of labor force entry in the two samples, caused by accelerated schooling of the more able. Whereas comparisons of earnings among individuals who differ only in amount of schooling should properly be adjusted for years of market experience,[17] comparisons among individuals differing only in ability should not. One of the major routes for collecting the return to high ability would appear to be early entry to the labor force, since earnings in the early years of the life cycle have a substantial weight in the calculation of present value of income streams.

[16] The importance of the quality of schooling can be gauged by the effect of tuition on earnings within the Terman sample, as discussed in Section II above. The importance of background variables for the Terman sample is discussed in my earlier paper.

### IV. Summary

The purpose of the present paper has been to argue that years of schooling do not adequately characterize the inputs

[17] See Mincer's (1974) development of the problem in *Schooling, Experience and Earnings.*

APPENDIX

Mean Incomes of Terman Sample and Census by Age and Education: 1940, 1950, 1960

| | High School 4 Years | | | College 1–3 Years | | | College 4 Years | | |
|---|---|---|---|---|---|---|---|---|---|
| Age | 1940 | 1950 | 1960 | 1940 | 1950 | 1960 | 1940 | 1950 | 1960 |
| | | | | Terman Sample[a] | | | | | |
| 22–24 | 3192 | | | 3380 | | | 3889 | | |
| 25–29 | 3655 | | | 3973 | | | 3736 | 4833 | |
| 30–34 | 3629 | 6200 | | 3495 | 10528 | | 4306 | 7689 | |
| 35–44 | 4278 | 10535 | 7500 | 3277 | 9259 | 14708 | 4981 | 10193 | 15520 |
| 45–54 | | 7375 | 12869 | | 5925 | 13954 | | 10808 | 16572 |
| 55–64 | | | 7800 | | | 9368 | | | 16462 |
| 65+ | | | | | | | | | |
| | | | | Census | | | | | |
| 22–24 | 869[b] | [c] | [d] | 936[b] | [c] | [d] | 1036[b] | [c] | [d] |
| 25–29 | 1210 | | | 1400 | | | 1645 | | |
| 30–34 | 1559 | 3488 | 5594 | 1804 | 3923 | 6059 | 2299 | 4695 | 7196 |
| 35–44 | 1987 | 3888 | 6606 | 2269 | 4602 | 8127 | 3160 | 6039 | 11118 |
| 45–54 | | 4296 | 6875 | | 4944 | 8801 | | 6578 | 13804 |
| 55–64 | | 4108 | 7004 | | 4500 | 8829 | | 6188 | 13638 |
| 65+ | | | | | | | | | |

*Note:* An equal distribution within income categories was assumed in calculating means.

[a] Calculated from the Terman sample.

[b] U.S. Bureau of the Census (1940), Table 33, p. 177. Data are weighted averages of 1939 wage and salary income classes reported by age and education for native white males with income and nonwage income less than $50 in the western region.

[c] U.S. Bureau of the Census (1950), Table 12, pp. 5B–116–17. Data are weighted averages of 1949 wage and salary income classes reported by age and education for males with income in North and West regions.

[d] U.S. Bureau of the Census (1960), Table 2, p. 197. Data are weighted averages of 1959 earnings classes reported by age and education of white males with earnings in experienced civilian labor force in North and West regions.

or the outputs of the process of investing in human capital. In particular, it has been demonstrated that the rate of return on schooling investments cannot be identified from the linear regression of *log* incomes on years of schooling, even given the knowledge of the average ratio of expenditures to full-time earnings. Empirically, I have shown that this ratio, $K$, is positively related to years of schooling, and that consequently rates of return which do not account for this correlation are biased upward. Lastly it has been shown that greater intensity of schooling investments may allow high ability students to enter the labor force earlier than their peers of average ability, and that the consequent increase in discounted value of lifetime earnings may be a major route of collecting the returns to high ability.

## REFERENCES


L. W. **Anderson**, "Student Involvement in Learning and School Achievement," paper presented to the American Educational Research Assn., Chicago, Apr. 15–19, 1974.

G. S. **Becker**, *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education,* New York 1964.

———, "Human Capital and the Personal Distribution of Income: An Analytical Approach," Inst. Publ. Admin., Univ. Michigan 1967.

——— and B. R. **Chiswick**, "Education and the Distribution of Earnings," *Amer. Econ. Rev. Proc.*, May 1966, *56*, 358–69.

Y. **Ben-Porath**, "The Production of Human Capital Over Time," in W. L. Hansen, ed., *Education, Income, and Human Capital,* New York 1970.

B. **Chiswick**, "Income Inequality and Schooling: A Cross-Sectional Study," mimeo., Nat. Bur. Econ. Res., Aug. 1972.

Z. **Griliches** and W. M. **Mason**, "Education, Income, and Ability," *J. Polit. Econ.*, May/June 1972, suppl., *80*, S74–S104.

W. J. **Haley**, "Human Capital: The Choice Between Investment and Income," *Amer. Econ. Rev.*, Dec. 1973, *63*, 929–44.

G. **Hanoch**, "An Economic Analysis of Earnings and Schooling," *J. Hum. Resources*, Summer 1967, *2*, 310–29.

J. **Hause**, "Earnings Profile: Ability and Schooling," *J. Polit. Econ.*, May/June 1972, suppl., *80*, S108–38.

T. **Johnson**, "Returns from Investment in Human Capital," *Amer. Econ. Rev.*, Sept. 1970, *60*, 546–60.

A. **Leibowitz**, "Family Investments in Human Capital," *J. Polit. Econ.*, Mar./Apr. 1974, suppl., *82*, S111–31.

L. A. **Lillard**, "Human Capital Life Cycle of Earnings Models: A Specific Solution and Estimation," Nat. Bur. Econ. Res. working pap. no. 4, July 1973.

J. **Mincer**, *Schooling, Experience and Earnings,* New York 1974.

——— and S. **Polachek**, "Family Investments in Human Capital: Earnings of Women," *J. Polit. Econ.*, Mar./Apr. 1974, suppl., *82*, S76–S108.

L. C. **Solmon**, "The Definition and Impact of College Quality," *Explor. Econ. Res.*, Fall 1975, *2*, 537–87.

P. **Taubman**, "Schooling, Ability, Non-Pecuniary Rewards, Socioeconomic background . . . ," Nat. Bur. Econ. Res. working pap. no. 17, Nov. 1973.

L. M. **Terman**, "The Promise of Youth," in *Genetic Studies of Genius,* Vol. III, Stanford 1930.

P. **Wachtel**, "The Returns to Investment in Higher Education: Another View," in F. T. Juster, ed., *Education, Income, and Human Behavior,* New York 1975.

D. **Wallace** and L. A. **Ihnen**, "Full Time Schooling in Life Cycle Models of Human Capital Accumulation," *J. Polit. Econ.*, Feb. 1975, *83*, 137–55.

**U.S. Bureau of the Census**, *1940 Census of Population,* Special Reports, Part 36, *Education,* Washington 1947.

———, *1950 Census of Population,* Vol. IV, Special Reports, Part 5-B, *Education,* Washington 1953.

———, *1960 Census of Population,* Subject Report PC (2)-7B, *Occupation by Earnings and Education,* Washington 1963.

# International Trade, Factor Market Distortions, and the Optimal Dynamic Subsidy

*By* HARVEY E. LAPAN*

International trade theorists have long favored free trade, arguing that the only proper basis for a tariff is in response to some monopoly power in international trade. Pragmatic politicians, on the other hand, wary of the wrath of their constituents, have often resorted to the use of tariffs, arguing that the protection afforded by tariffs was necessary to prevent unemployment. Recent papers on the theory of domestic distortions and optimal policy interventions can be interpreted as an attempt to reconcile these two divergent views. For example, if factors are immobile and if distortions exist in the factor markets (due to factor price rigidities), then it is argued that the optimal policy intervention is not a tariff, which destroys the equality between the marginal rate of substitution (*MRS*) and foreign rate of transformation (*FRT*), but rather a wage subsidy. This policy prescription, which recognizes the pragmatic difficulties politicians must face, offers a feasible remedy that is proved superior to the use of tariffs.

However, one can argue that the theory of optimal policy intervention goes a bit too far. While it may well be true that factors are not instantaneously mobile, they certainly do become mobile through time; capital depreciates, new workers enter the labor force, and factories can be moved intertemporally. The policies that eliminate unemployment at the same time destroy the incentives for resource reallocation (particularly if the institutional constraint stipulates that factor prices must be the same in all sectors). Thus, while the standard wage subsidy that would promote full employment may be efficient in the short run, it may prove inefficient in a dynamic setting.

Nevertheless, this does not imply that the efficient long-run policy entails no intervention; rather, it is likely that it will call for some wage subsidy which is smaller than the static optimum subsidy. It is the purpose of this paper to discuss how this dynamic optimum subsidy is determined and to specify its implications for the economic system.

## I. The Problem

For the analysis, I shall use a standard two-sector, one-factor trade model.[1] Let the two commodities be $M$ and $C$, and let $N_m$ and $N_c$ represent the labor employed in each sector. Then:

$$(1) \qquad M = F_m(N_m), \qquad C = F_c(N_c)$$
$$F_i' > 0, \quad F_i'' < 0; \qquad i = c, m$$

Furthermore, choose $C$ as the numeraire and let $P$ represent the relative price of $M$.

In addition, assume that the number of potential workers in each sector ($L_i$) is fixed at any moment, and that the total

* Associate professor of economics, Iowa State University. I am indebted to Gordon Rausser and Walter Enders for their valuable comments and suggestions.

[1] If desired, it can be assumed that other factors are used in each sector, but that these are fixed for the period under study. John Harris and Michael Todaro use this model.

supply of available workers is fixed for the period in question:

$$(2) \qquad N_i \leq L_i, \qquad i = c, m$$
$$L_c + L_m = \bar{L}$$

Institutionally, it is assumed that firms behave competitively in factor (and product) markets, so that workers are paid their marginal value product (and there cannot be excess demand for factors in either sector). However, it is also assumed that because of distortions in the factor market, the wage rate must be the same in each sector.[2] Thus, in the absence of any wage subsidies, we have:

$$(3) \qquad F_c'(N_c) = PF_m'(N_m)$$
$$N_c = L_c \quad \text{or} \quad N_m = L_m$$

The above three equations constitute a long-run equilibrium if commodity markets clear and if full employment occurs in each sector.

To simplify the analysis, let us assume that the economy in question is open and "small," so that it can trade at unchanging terms of trade.[3] Assuming the economy is in long-run equilibrium it will be producing at some point (such as $B$ in Figure 1) on the long-run production possibility frontier and will consume somewhere on the price line tangent to the production possibility frontier at $B$.[4]

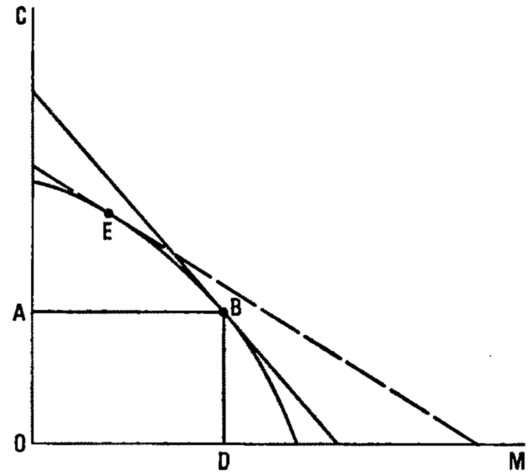



Figure 1

Now suppose that due to external conditions, the terms of trade shift; for example, assume the country was exporting $C$ and let its terms of trade improve ($P$ falls). If factors are immobile, the short-run production possibility frontier is given by $ABD$; and if no distortions exist in the factor markets, the country will continue to produce at $B$ and will benefit from the improved terms of trade. However, if factors are immobile and if factor prices must be the same in each sector, an excess supply of labor will develop in $M$ and output will occur somewhere along the open segment $AB$. Thus, it is conceivable that the improved terms of trade might leave the country worse off.

As has been demonstrated in the literature,[5] the optimal policy under these circumstances is a wage subsidy to $M$ that restores production at point $B$. Thus, if we let $\bar{P}$ be the old price at which resources were fully employed, and $P^*$ the new world price ($P^* < \bar{P}$), and if $S$ represents the optimum wage subsidy to sector $M$, then:

[2] A common alternative assumption is that "real" wages are rigid downward; the implications of this assumption depend on the definition of real wages. If they are defined in terms of $C$, then (assuming we start from full employment) a fall in $P$ will create unemployment, but a rise in $P$ will have no adverse effect on employment; opposite results hold if real wages are defined in terms of $M$. Finally, if they are defined in terms of some weighted price index, any change in $P$ will cause unemployment; this latter assumption gives results that are qualitatively the same as assuming wages must be equalized across sectors.

[3] This assumption, which is commonly used (see Jagdish Bhagwati and T. N. Srinivasan) allows us to ignore the utility function and it implies the optimum tariff is zero. The results would not be significantly altered by dropping this assumption.

[4] The long-run production possibility frontier represents the efficient production locus when labor is assumed perfectly mobile.

[5] For example, see the articles by Harry Johnson and Stephen Magee.

$$(4)\quad F_c'(L_c) = \bar{P}F_m'(L_m) = P^*F_m'(L_m)/(1-S)$$
$$S = [(\bar{P} - P^*)/\bar{P}] > 0$$

This subsidy restores full employment and enables the country to benefit from its improved terms of trade.

However, this equilibrium is still suboptimum in a long-run context since the change in the terms of trade moves the long-run optimum production point from $B$ to $E$ (in Figure 1). The obvious question is what determines intertemporal factor migration and what is the best policy to be pursued in a dynamic context? If, for example, factor migration between sectors is determined by economic factors, such as different wages or unemployment rates in each sector, then the static optimum subsidy ($S$) destroys any economic incentives for migration since, by assumption, wages are equalized between sectors while, by choice, unemployment is reduced to zero in each sector. Thus, while $S$ may be efficient in a short-run context, it may not be optimal when considered in a long-run context.[6]

Clearly, the government may have other tools at its disposal that will enable it to maintain full employment in the short run while encouraging migration in the direction of long-run equilibrium. One policy tool, suggested in Harris and Todaro, would be simply to force (or restrict) migration in a particular direction while maintaining the optimal subsidy $S$ needed to guarantee full employment of resources if there were factor market distortions (naturally, $S$ changes through time as labor migrates). However, this particular policy is most likely to be opposed because it infringes on personal liberty (and it is not likely to be very efficient if all workers are not identical).

The economic tools the government should use to maintain full employment and encourage migration depend upon the causes of migration and the institutional constraints.[7] If migration rates between sectors depend upon differences in real incomes between the sectors, the government should attempt to enlarge these differences, while at the same time pursuing a policy that maintains full employment in all sectors. For example, if we do *not* assume that wages must be equalized across sectors, then the government could apply large wage subsidies to sector $C$ (where labor is more productive) to encourage migration toward $C$, while at the same time use the minimum wage subsidy necessary to promote full employment in $M$. Even if wages *must* be equalized across sectors, the government could encourage migration to $C$ by imposing a lump sum tax on workers in $M$ (or a lump sum transfer to workers in $C$), while at the same time subsidizing wages in $M$ in order to promote full employment.

However, the above analysis is subject to criticism on several grounds. It assumes that somehow labor (effectively) demands a minimum real wage but apparently cannot effectively demand a minimum real standard of living. Specifically, the above policy can work because it assumes that a wage subsidy can be applied to sector $M$ to meet minimum wage demands and promote full employment, while (some of) the income from this wage subsidy can be removed by lump sum taxation (to promote real income differences between the sectors) without altering the wage demands of laborers in this sector.

Moreover, it can be argued that the

[6] Even if no distortions exist in factor markets, so that full employment (at $B$) occurs without subsidization, it does not follow that this solution is optimal in a long-run context when we are interested in encouraging migration between sectors.

[7] In the following analysis it is assumed that labor's marginal value product is higher in $C$, so that the government wishes to encourage migration toward $C$ while maintaining full employment in $M$ (by means of an optimum subsidy).

economic efficiency of the wage subsidy to $C$ (to promote migration) depends crucially upon the assumption that the return to labor is a pure economic rent. If each individual's labor supply decision were responsive to the wage rate, then the subsidy in $C$ would create inefficiencies by causing a discrepancy between the marginal value product of labor and that worker's marginal rate of substitution between consumption and leisure. Thus, while the wage subsidy (to the more efficient sector) attempts to promote migration, and hence dynamic efficiency, it produces static inefficiency.

Given the preceding discussion, we are skeptical about the ability of the government to costlessly reallocate resources in the economy. Thus for the remainder of the paper we explicitly assume this reallocation *cannot* be achieved costlessly, and that it is the unemployment rate (in $M$) that determines the rate of out-migration. One might rationalize this assumption by arguing that workers are risk averse and hence are unwilling to move to an unfamiliar situation (a new industry) if there is a reasonable chance of finding a job in their own sector. In addition, if it is *assumed* that wages must be the same in all sectors, the assumption that unemployment rates determine migration is equivalent to the assertion that migration is determined by the differences in expected wages between the sectors.[8]

In the analysis that follows, no explicit assumption will be made about wage rigidity in either sector; rather, the key relation will be the effect of unemployment rates on migration. The central planner's task is to determine the time path of the unemployment rate that balances the costs of unemployment (due to lost output) against the gains from migration (due to a more efficient labor allocation). However, given the institutional constraints (shown in fn. 8), the optimal wage subsidy can be inferred from the time path of the unemployment rate.

In the following section we develop the formal model and discuss the properties of an optimal solution; attention is focused on determination of the optimal unemployment rate. The principal conclusion of the section is that for long planning horizons (or small discount rates), some unemployment is desirable during early stages of the plan. Since this implies that the static subsidy is not in general dynamically efficient, the properties of the optimal dynamic subsidy are discussed in Section III.

## II. Optimal Labor Transfer

Suppose the economy is out of long-run equilibrium in the sense that labor's marginal value product is larger at full employment in sector $C$ than in $M$.[9] Furthermore, assume that labor is not instantaneously mobile, but that it moves through time in response to unemployment. We seek to determine the optimum time path of unemployment, assuming that the central planner wishes to maximize the present discounted value of the stream of income over the interval (0, T).[10,11]

Let production and factor endowments be as described in equations (1) and (2).

[8] For further discussion of the role of expected wages, see Harris and Todaro. The particular institutional assumptions we make for the analysis of Sections II and III are: (i) Wages to workers must be the same in each sector and there cannot be an excess demand for labor in either sector: $W_c = W_m = \text{Max}[F_c'(L_c), PF_m'(L_m)]$; (ii) Migration between sectors occurs due to differences in the expected wage rate, and hence responds to differences in the unemployment rate between sectors; (iii) The government controls the unemployment rate through wage subsidies. Consequently, there is a tradeoff between current output and future output in that large subsidies that promote full employment do so at the expense of migration that would ultimately raise the value of output.

[9] The difference in marginal value products may have arisen because of some exogenous change in the terms of trade. Also, the analysis could readily be generalized to handle the case in which labor's marginal value product is larger in $M$.

[10] Given the assumption the economy is open and "small," maximization of the present value of the in-

Since we wish to encourage migration towards $C$, it will always be optimal to maintain full employment there. Thus, labor migration towards $C$ can be postulated as a function of the unemployment rate ($u$) in $M$:

$$(5) \quad \frac{dL_c}{dt} = \frac{-dL_m}{dt} = \Phi(u)L_m, \quad u \in [0, 1]$$

$$\Phi(0) = 0; \quad \Phi' > 0, \ \Phi'' \leq 0$$

From the definitions and the initial conditions:

$$(6) \quad N_c = L_c, \quad N_m = L_m(1 - u)$$

$$(7) \quad F_c'(L_c(0)) > PF_m'(L_m(0))$$

Given (1), (2), (5), (6), and (7), we seek to maximize:

$$(8) \quad V = \int_0^T [F_c(L_c) + PF_m(N_m)]e^{-rt}dt$$

The central planner must determine the time path of the unemployment rate that maximizes the present discounted value of $GNP$ over the interval. Since the costs of unemployment depend upon labor's marginal value product in $M$ (and are immediate), while the benefits due to migration depend upon the differences in marginal value products between sectors and are deferred, it follows that the optimal policy (at each period) depends upon the current labor allocation, the discount rate, and the amount of time left in the planning horizon. Intuitively, we expect that unemployment is undesirable near the end of the planning horizon, but (relatively) more desirable during the early stages of the plan.[12]

Formally, we seek to determine the time path of the unemployment rate $u^*(t)$ that maximizes (8) subject to (1), (2), (5), (6), and (7). The problem as stated is an optimal control problem in one state variable $L_c(t)$ and one control variable ($u$). The Hamiltonian is:[13]

$$(9) \quad H = [F_c(L_c) + PF_m(N_m)]e^{-rt} + \lambda[L_m\Phi(u)]$$

In (9), $\lambda(t)$ represents the value, discounted to time 0, of transferring a worker at time t to sector $C$. It is convenient to define $q(t)$ to represent the value at time t of such a transfer:

$$(10) \quad q(t) = \lambda(t)e^{rt}$$

Optimizing the Hamiltonian over $u$ yields:[14]

$$(11) \quad H_u = L_m e^{-rt}[q\Phi'(u) - PF_m'(N_m)] \leq 0 \qquad u \in [0, 1]$$

In addition, for an optimum path we derive the canonical equations and the transversality condition:

$$(12) \quad \frac{dq}{dt} = [r + \Phi(u)]q - [F_c' - P(1 - u)F_m']$$

come stream is equivalent to utility maximization if either: (i) the marginal utility of income is constant and $r$ is the rate of time preference; or (ii) the country can borrow or lend on world capital markets at the rate $r$. In the latter case, the planner must still determine the optimum consumption stream. Our results would be qualitatively the same if we worked directly with a utility function.

[11] If labor supply decisions are made under utility maximization, then it is necessary to choose an objective function that contains leisure as a variable. Moreover, we could not use a community indifference curve for the analysis since expected wages would differ across sectors; rather, we would need a Bergson-type Social Welfare Function. We assume labor supply is exogenous because this is in the spirit of the previous work on factor market distortions.

[12] The unemployment rate plays a role similar to that of savings in the one-sector growth models. Savings is undesirable during the latter stages of those models unless a value is placed on the terminal capital stock.

[13] The model (1)–(8) can be used to analyze optimal labor transfer even if no factor price rigidities exist (or if it is not unemployment that induces mi ration), provided this transfer is costly due to time lost from work.

[14] We assume $F_m'(0) = \infty$; thus some employment is always desirable in $M$. Also, we assume $\Phi'(0)$ is finite; if it is not, some unemployment will always be optimal (if marginal value products differ), and the static subsidy will always be inefficient.

$$\frac{dL_c}{dt} = L_m\Phi(u) \tag{13}$$

$$q(\mathrm{T}) = 0, \qquad L_c(0) \text{ given} \tag{14}$$

By the assumptions on $F_i''$ and $\Phi''$, any path that satisfies (11)–(14) constitutes an optimum solution to the problem. In order to characterize the properties of this solution, it is helpful to construct the phase diagram for the system of differential equations in $(L_c, q)$ space.[15]

First, consider (11) which serves to determine (for given $q$, $L_c$) the optimal unemployment rate; it states that if some unemployment is desirable, then the social value attributable to an additional unemployed worker (due to induced migration and equal to $q\Phi'(u)$) should equal the value of output lost due to his unemployment. However, if $H_u < 0$ for all $u$, then full employment is desirable. Thus, the boundary between full employment and some unemployment is defined by the locus such that $H_u = 0$ for $u = 0$:

$$g(q, L_c) = [q\Phi'(0) - PF_m'(L_m)] = 0 \tag{15}$$

As shown in the Appendix, this curve (and all iso-unemployment curves) is positively sloped in $(L_c, q)$ space and is asymptotic to the line $L_c = \bar{L}$. Larger values of $q$ are associated with higher unemployment rates (given $L_c$); thus, points above the $g(q, L_c)$ locus correspond to $u > 0$, $(dL_c/dt) > 0$, whereas points below it correspond to $u = 0$, $(dL_c/dt) = 0$. The curve $AB$ in Figure 2 depicts this locus.

Next, consider how $q(\mathrm{t})$ changes along

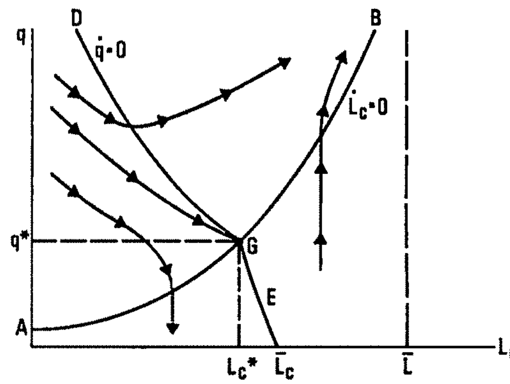



FIGURE 2

an optimum path. Since $q(\mathrm{t})$ is the value of labor transfer at time t, it must fall through time since: (i) the differences in marginal value products between sectors is a nonincreasing function of time, and (ii) as time passes, less time (for finite T) remains to reap the benefits of further labor transfers. To determine the optimal region in $(L_c, q)$ space it is useful to sketch the curve $(dq/dt) = 0$. From (11), $u$ can be written as an implicit function of $L_c$ and $q$; substitution in (12) yields:

$$\begin{aligned} dq/dt = h(q, L_c) = {} & (r + \Phi)q \\ & - [F_c' - P(1-u)F_m'] = 0 \end{aligned} \tag{16}$$

In the Appendix it is shown this locus is negatively sloped and intersects the $q = 0$ axis at $\bar{L}_c$, the labor allocation that equalizes marginal value products across sectors.[16] Points above the locus correspond to $(dq/dt) > 0$, whereas those below correspond to $(dq/dt) < 0$; this locus is sketched in Figure 2 as the curve $DE$.

Given this phase diagram, $q(0)$ must be chosen to satisfy (14), the transversality condition; from the previous arguments, $(dq/dt) < 0$ and thus $q(0)$ lies below $DE$. Once $q(0)$ is chosen, the unemployment rate (and the subsidy rate) can be de-

[15] In the following, we assume $r > 0$; even if the social discount rate is zero, it is appropriate to choose $r > 0$ if borrowing and lending can occur (at a positive rate) on world capital markets. When it is appropriate, we shall indicate the implications of assuming $r = 0$. Note that if $r = 0$, (8) will not converge as $\mathrm{T} \to \infty$. However, an equivalent problem would be to minimize the deviation of actual output from maximum output: $[F_c(\bar{L}_c) + PF_m(\bar{L}_m) - F_c(L_c) - PF_m(N_m)]$, where $F_c'(\bar{L}_c) = PF_m'(\bar{L}_m)$. Since this integral converges, and since the two problems give equivalent conditions, we can proceed with the problem as stated.

[16] For $r = 0$, the locus $(dq/dt) = 0$ intersects the $g(q, L_c) = 0$ locus at $\bar{L}_c$, and it becomes vertical at this point, intersecting the $q = 0$ locus at $\bar{L}_c$. However, for $u > 0$, the locus is negatively sloped.

duced from Figure 2. To further clarify the nature of the solution, define $(L_c^*, q^*)$ as the point ($G$ in Figure 2) at which (15) and (16) intersect. Solving these equations simultaneously yields:

$$(17)\quad F_c'(L_c^*) - PF_m'(L_m^*) = PF_m'(r/\Phi'(0)) > 0;$$
$$q^* = PF_m'/\Phi'(0), \quad L_c^* < \bar{L}_c \quad \text{for } r > 0$$

If $L_c(0) \geq L_c^*$, then full employment is always optimal, regardless of the length of the planning horizon.[17] Since $L_c^*$ is a decreasing function of $r$, it follows that for relatively large discount rates or a relatively small misallocation of labor, it does not pay to incur unemployment, and thus the static subsidy is dynamically efficient.[18]

However, if $L_c(0) < L_c^*$, the optimal solution depends on T. For this case, it can be shown that the solution to the differential equations is a saddlepoint, and that the unique path which converges to $(L_c^*, q^*)$ is the turnpike—the optimum solution for the infinite time horizon problem.[19] The infinite time horizon problem is characterized by decreasing $q(\mathrm{t})$ and increasing $L_c(\mathrm{t})$, and thus falling unemployment rates. Note that even in this case, it does not pay to equalize labor's marginal value product across sectors.[20]

For any finite T, it is clear that the choice of $q(0)$ lies below the turnpike—and thus less labor will be transferred during the planning period. Moreover, it is readily seen that the shorter the time horizon, the less desirable unemployment becomes, so that the initial unemployment rate (and $q(0)$) decreases, and the total labor transfer decreases, as T decreases. Since $q(\mathrm{t})$ falls along an optimum path, the unemployment rate falls (for $u>0$) along this path.

Moreover, the terminal stage of any finite horizon path must be characterized by full employment. To show this, assume an optimal path has been followed during the interval $(0, \mathrm{T}-\tau)$, and let $L_c(\mathrm{T}-\tau)$ represent the corresponding labor allocation. The cost of incurring unemployment during the interval $(dt)$ is $A$:

$$(18)\qquad A \simeq PL_m u F_m' d\mathrm{t}$$

The increased flow of output due to this labor transfer, assuming full employment subsequently, is:

$$(19)\qquad B = (F_c' - PF_m')\Phi(u)L_m d\mathrm{t}$$
where $dL_c = \Phi(u)L_m d\mathrm{t}$

Discounting future benefits and integrating over the remaining time yields:

$$(20)\quad B' \simeq (F_c' - PF_m')\Phi(u)L_m[(1-e^{-r\tau})/r]dt$$

If $(B'-A)$ is nonpositive for all $u$, then full employment is optimal for the whole interval. Thus, $u=0$ if:[21]

$$(21)\quad e^{-r\tau} \geq [F_c' - PF_m'(1 + [r/\Phi'(0)])] / (F_c' - PF_m') \qquad r > 0$$

If the numerator on the right-hand side is nonpositive $(L_c \geq L_c^*)$, then full employment is optimal, regardless of $\tau$. Even for

[17] By assumption, $L_c(0) < \bar{L}_c$; for $r=0$, $L_c^* = \bar{L}_c > L_c(0)$. For $r>0$, $L_c(0) \geq L_c^*$, $u \equiv 0$ is optimal; $q(0)$ is chosen so that $q(\mathrm{T})=0$. If the horizon is infinite, $q(0)$ is chosen so that $(dq/d\mathrm{t})=0$.

[18] This is analogous to the results of the one-sector growth model when there is no population growth or capital depreciation. In this model, steady-state per capita consumption is maximized at the capital-labor ratio $(k^*)$ such that $f'(k^*)=0$. However, the modified golden rule determines $\hat{k}$ such that $f'(\hat{k})=r$. If $k^* > k(0) > \hat{k}$, then the optimal policy is to do nothing; if $k(0) < \hat{k}$, some savings will be optimal for a sufficiently long planning horizon.

[19] Naturally, if T is unbounded, the transversality condition given by (14) must be suitably modified.

[20] For $r=0$, the marginal value products are equalized asymptotically. It can be shown that for the infinite horizon case, $(L_c^*, q^*)$ is not reached in finite time; thus some unemployment is always desirable, though $u(\mathrm{t}) \to 0$. The proof is omitted to save space.

[21] If $r=0$, (21) becomes $\tau \leq PF_m'/[\Phi'(0)(F_c' - PF_m')]$. Thus, for $r=0$, large $\tau$ implies $u>0$. Nevertheless, it is clear that it does not pay to equalize labor's marginal value product for any finite horizon.

$L_c(0) < L_c^*$, full employment is optimal for small $r$ (since the right-hand side of (21) is less than one). Thus, for short planning periods, and for the terminal stage of any finite plan, it is optimal to maintain full employment.

From (21), it is apparent that given $r$ and $L_c(0) < L_c^*$, there exists a $T^*$ such that for $T < T^*$, $u(t) \equiv 0$; for $T > T^*$, the initial stages of the optimal solution are characterized by positive but decreasing unemployment rates, and the terminal stage by full employment. As noted earlier, $L_c(T)$ is an increasing function of T;[22] thus (21) implies that the length of the period of full employment increases as T increases.

To recapitulate, I have shown that (for $r>0$) it is never optimal to fully reallocate labor between the sectors. Also, we have seen full employment is always optimal either if the discount rate is large or the planning horizon is short; in these cases, the static subsidy that maintains full employment is optimal. However, if $r$ is not large (or initial misallocations of labor are large) and if the planning horizon is sufficiently long, unemployment is desirable during the initial stages of the plan; for these cases, the static subsidy is inefficient. Therefore, I conclude that the static subsidy represents myopic behavior because it discourages resource reallocation.

## III. The Economic Properties of the Optimal Path

The preceding section discussed the qualitative properties of an optimal solution; we now derive some of the quantitative properties of this solution. Specifically, I shall discuss how: (i) *GNP*, (ii) employment in *M*, and (iii) the optimal subsidy change through time.

If full employment is always optimal, then all economic variables are constant, and the static subsidy is dynamically efficient. However, we have argued that this solution represents myopic behavior; in the following analysis it is assumed that some unemployment is optimal for the initial stage of the plan.

First, consider the marginal value product of labor in each sector. Given the assumptions, $F_c'(L_c) > PF_m'(L_m)$; however, since there is unemployment in *M*, it does not immediately follow that $F_c'(L_c) > PF_m'(N_m)$. From (11) and (12) we have, for $u>0$:

$$(22)\quad dq/dt = [r + \Phi - u\Phi']q - [F_c' - PF_m'(N_m)] < 0$$

since $q(t)$ falls along the optimum path. But $[\Phi - u\Phi'] \geq 0$ for all $u$, since $\Phi'' \leq 0$; thus (22) implies $F_c'(L_c) > PF_m'(N_m)$ along the optimum path.

Consider how *GNP* changes through time; I have shown that $L_c(t)$ increases and $u(t)$ decreases through time ($u>0$). From (22) it immediately follows that *GNP* is increasing through time since total employment is rising and labor is being reallocated to its more productive use.

Less apparent is how total employment in *M* changes through time—$u(t)$ is falling, as is $L_m(t)$, so the net result could be ambiguous. Without further assumptions on $\Phi''$ or $F_i''$, it does not appear possible to describe how $N_m$ changes through time. However, if $\Phi''=0$, the answer is apparent. From (11), for $u>0$:

$$(11')\quad \Phi'(u)\cdot q = \Phi'(0)\cdot q = PF_m'(N_m)$$

$$\frac{dq}{dt} < 0$$

Since $q$ decreases if $\Phi''=0$, then $PF_m'(N_m)$ must also decrease through time. Thus, for $\Phi''=0$, total employment in *M* rises along the optimum path ($u>0$).

Furthermore, this same result holds during the final portion (of the period of unemployment) for any finite horizon path.

[22] This can be demonstrated by comparing the optimal paths for the horizons T and (T+$d$T), respectively.

During this portion $u$ must be small, and since $dq/dt$ does not tend to zero as $u$ tends to zero (for finite T), it follows that $\Phi'(u)\cdot q$ and hence $PF'_m(N_m)$ must be falling.[23] Therefore, towards the end of the period of unemployment, $N_m$ must rise.

Finally, we consider the behavior of the optimum dynamic subsidy $S^*(t)$, assuming wages must be equalized across sectors. Given this assumption, the static subsidy needed to maintain full employment is:[24]

$$(23)\quad S(t) = 1 - [PF'_m(L_m)/F'_c(L_c)] > 0 \qquad L_c < \bar{L}_c$$

The dynamic subsidy depends upon the unemployment rate:

$$(24)\quad S^*(t) = 1 - [PF'_m(N_m)/F'_c(L_c)] \qquad S^* < S, \quad u > 0$$

Note that (22) implies $S^*(t)>0$. Since $S^*<S$ for $u>0$, and $S^*=S$ for $u=0$, the gap between these subsidies must eventually be eliminated for any finite horizon. However, without further assumptions, it does not appear possible to conclude that the gap monotonically decreases through time.

If we compare the fraction of wages paid by employers under each subsidy, we can determine how this ratio changes through time if the production function in $M$ is isoelastic. Specifically, let:

$$(25)\quad F_m(N_m) = N_m^{\alpha}, \qquad \alpha \in (0, 1)$$

From (23) and (24) we have, given (25):

$$(26)\quad [(1 - S^*)/(1 - S)] = [F'_m(N_m)/F'_m(L_m)] = (1 - u)^{\alpha-1} > 1, \qquad u > 0$$

Since $du/dt<0$, $(1-S^*)/(1-S)$ decreases through time, approaching one. Therefore, the ratio of the percent of wages paid by employers in the dynamic case to the static case decreases through time. However, even this does not imply $(S-S^*)$ monotonically decreases through time; without (25) it does not appear possible to reach specific conclusions.

The final issue to be resolved concerns the time path of the optimum subsidy. From (24) it is apparent that if $N_m(t)$ decreases through time, then $S^*$ also decreases. However, we have shown that $dN_m/dt>0$ for $\Phi''=0$ and for latter portions of the optimal path. Thus, the time path of $S^*(t)$ is not immediately apparent.

Further information can be obtained by considering the time derivative of (24). Letting $PF'_m=q\Phi'$ in (24) and differentiating yields:

$$(27)\quad \frac{dS^*}{dt} = \frac{-\Phi''q(du/dt)}{F'_c} - \frac{\Phi'(dq/dt)}{F'_c} + \frac{F''_c\Phi'q(dL_c/dt)}{(F'_c)^2}$$

While the sign of (27) is not known for all t, it is clear that $dS^*/dt>0$ near the end of the period of unemployment (for finite T). Therefore, the optimum subsidy must be rising during the final stages of unemployment.[25]

Rewriting (27), assuming $\Phi''=0$, yields:

$$(27')\quad dS^*/dt = -[q\Phi'/F'_c]\left[r - \frac{S^*\Phi'}{(1-S^*)} + \frac{\beta\Phi L_m}{L_c}\right]$$

where $\beta\equiv-(L_cF''_c/F'_c)>0$. Assuming $\beta$ is constant, differentiating (27′) with respect to t and evaluating it at $[dS^*/dt]=0$ yields:

[23] For any $u$, $\Phi(u)=\Phi(0)+\Phi'(0)\cdot u+\Phi''(\hat{u})\cdot u^2/2$, for some $\hat{u}\in[0, u]$. Thus, if $u$ is small, we can approximate: $\Phi(u)=\Phi'(0)\cdot u$, since $\Phi(0)=0$.

[24] Naturally, the static subsidy changes as the labor allocation changes; thus, $S(t)$ decreases as $L_c(t)$ increases.

[25] This result holds for finite T since $u$ is a continuous function of time, and thus $(dL_c/dt)$ and $(du/dt)$ tend to zero as $u\to0$. However, $(dq/dt)<0$ and $\Phi'>0$ implies $(dS^*/dt)>0$ as $u\to0$.

$$(28)\quad \left[\frac{d^2S^*}{dt^2}\right]_{dS^*/dt=0} = -\left[\frac{q\Phi'}{F_c'}\right]\left[\frac{\beta\Phi' L_m(du/dt)}{L_c} - \frac{\beta\Phi^2\overline{L}L_m}{L_c^2}\right] > 0$$

The sign of (28) follows directly from the result $du/dt<0$ for $u>0$.

Equation (28) says that once $S^*(t)$ starts to increase, it must continue to do so. It follows that for any finite horizon path, if ever $[dS^*/dt]<0$, this must occur early in the plan. For infinite horizon paths, (28) is even more informative; if $r=0$, $S^*(t)\to 0$ asymptotically. Since $S^*(t)>0$ for $u>0$, it follows that the optimal subsidy falls monotonically for the infinite horizon, zero-discount case (given the constancy of $\beta$ and $\Phi'$).

Similar results hold for $r>0$. Let $\hat{S}$ be the optimum subsidy corresponding to the stationary point $(L_c^*, q^*)$; asymptotically $S^*(t)\to\hat{S}$ for the infinite horizon path $(L_c(0)<L_c^*)$. From (17) and (24):

$$(29)\quad \hat{S} = [r/(\Phi'+r)]$$
$$\hat{S}/(1-\hat{S}) = (r/\Phi')$$

From (27′) it is clear that $dS^*/dt<0$ for $S^*\leq\hat{S}$, $u>0$; but this implies that $S^*(t)$ tends to $\hat{S}$ from above, so that $S^*(t)$ decreases monotonically for the infinite horizon case.[26]

Equations (27′), (28), and (29) imply that for any finite T, $S^*(t)>\hat{S}$, since $S^*(t)$ rises during the terminal stages of the finite horizon path. Moreover, these equations imply that for large T, $S^*(t)$ must be decreasing during early stages of the plan. This result holds since, given $L_c(0)$, the larger is T, the larger is $u(0)$, and the smaller is $S^*(0)$; further, as T increases, these values approach the rates for the infinite horizon solution, and I have already shown $S^*(t)$ decreases monotonically

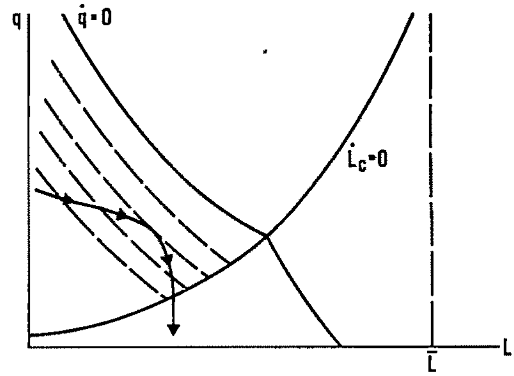



FIGURE 3

in that case. Thus, given $L_c(0)<L_c^*$, there exists a $\overline{T}$ such that for $T<\overline{T}$, $S^*(t)$ rises monotonically along the optimum path, whereas for $T>\overline{T}$, $S^*(t)$ first decreases, reaches a minimum (which exceeds $\hat{S}$), and then starts increasing (for $u>0$).

Thus, the time path of $S^*(t)$ for any finite horizon can be characterized as follows:

(i) for small T, $S^*$ is constant and equal to the static subsidy;
(ii) for intermediate values of T, $S^*(t)$ rises for $u>0$, and then remains constant for $u=0$;
(iii) for larger values of T, $S^*(t)$ initially decreases, reaches a minimum, then increases until $u=0$.[27]

Figures 3 and 4 depict these results (the dashed curves in Figure 3 represent iso-

[26] I have shown $S^*(t)>\hat{S}$; thus $S^*(t)$ approaches $\hat{S}$ monotonically from above since once $S^*$ starts to increase, it must continue to do so.

[27] For initially large misallocations of labor, it is likely that $dS^*/dt<0$ during early stages of the plan. To see this, rewrite (27′) as:

$$\frac{dS^*}{dt} = -\left(\frac{q\Phi'}{F_c'}\right)\left[r+\Phi'+\frac{\beta\Phi' uL_m}{L_c} - \frac{\Phi' F_c'}{PF_m'}\right)$$
$$= -\left(\frac{q\Phi'}{F_c'}\right)\left[r+\Phi'+\frac{\Phi'}{L_c}\left\{\beta uL_m - \left(\frac{L_cF_c'}{F_c}\right)\left(\frac{F_c}{PF_m'}\right)\right\}\right]$$

where $\Phi=\Phi'(0)\cdot u$ since $\Phi''=0$. Given T, let $L_c(0)\to 0$; then, if $F_c=L_c^{1-\beta}$, $F_c\to 0$ as $L_c\to 0$ and $(L_cF_c'/F_c)=(1-\beta)$; also $(uL_m)$ increases as $L_c(0)$ decreases and $PF_m'$ remains nonzero, given finite $\overline{L}$. Thus, there exists $\hat{L}_c$ such that for $L_c(0)<\hat{L}_c$, $(dS^*/dt)(0)<0$, indicating $S^*$ decreases during early stages of the plan.

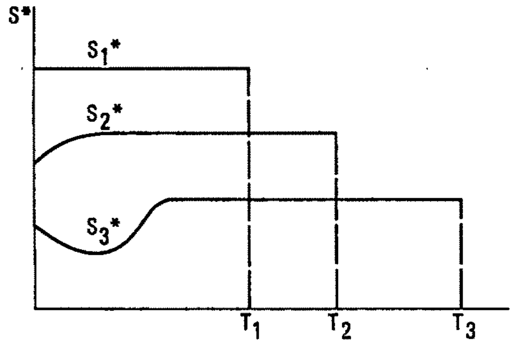



FIGURE 4

subsidy lines). Figure 4 is drawn under the assumption that $L_c(0)$ is constant; all that varies is the length of the planning horizon.[28] Finally, note that regardless of the assumptions on $\Phi''$ and $[L_cF_c''/F_c']$, $S^*(t)$ must increase during the final stages of unemployment for any finite horizon path.

## IV. Conclusion

We have seen that if resources cannot be transferred costlessly, the static subsidy is inefficient in a dynamic context unless the planning horizon is short, or discount rates are large. Thus in some sense the static subsidy represents myopic behavior. Moreover, I have shown how the optimum path can be determined and have characterized the properties of the path. In particular, we have seen that for long time horizons and low discount rates, it will always be optimal to have some unemployment initially; and the initial level of unemployment increases with the time horizon. Nevertheless, if wages must be equalized across sectors, some subsidy will always be needed; and we have discussed how this subsidy changes through time. Thus, a realistic policy must recognize that resources are not instantaneously mobile, but it must also recognize that too large a subsidy removes the incentives for intertemporal reallocation of resources.

[28] In Figure 4, the vertical broken lines are meant to indicate the end of the planning horizon, at which time the subsidy program ends.

## APPENDIX

To show iso-unemployment loci are positively sloped, consider (11), which defines an implicit function in $u$, $L_c$, and $q$. Differentiating (11) yields:

$$\text{(A1)} \quad \partial u/\partial q = -\Phi/[q\Phi'' + PL_mF_m''] > 0$$

$$\text{(A2)} \quad \partial u/\partial L_c = -[(1-u)PF_m''] \,/\, [q\Phi'' + PL_mF_m''] < 0$$

Thus, all iso-unemployment loci are positively sloped; also from (11) it is clear that as $L_c \rightarrow \overline{L}$, $q \rightarrow \infty$ for $H_u = 0$. From (A1), it is seen that points above the $g(q, L_c) = 0$ locus represent positive unemployment.

For the locus $h(q, L_c) = 0$, defined by (16), we find by differentiating:

$$\text{(A3)} \quad \partial h/\partial q = [r + \Phi] - PN_mF_m''(\partial u/\partial q) > 0$$

$$\text{(A4)} \quad \partial h/\partial L_c = -F_c'' - [P(1-u)^2 q\Phi''F_m''] \,/\, [q\Phi'' + PL_mF_m''] > 0$$

Equations (A1) and (A2) are used in deriving (A3) and (A4) since $u$ is defined as an implicit function of $L_c$ and $q$ by (11). Thus, the $(dq/dt) = 0$ locus is negatively sloped. However, for $r = 0$, $u = 0$, $(\partial h/\partial q) = 0$, indicating that the locus becomes vertical at the point where it crosses the $g(q, L_c) = 0$ locus. For $r > 0$, $u = 0$, it is readily seen from (11) that $(dq/dt) = 0$ at $q = 0$, $F_c' = PF_m'$. Finally, (A3) asserts that points above the $h(q, L_c) = 0$ locus correspond to $(dq/dt) > 0$.

## REFERENCES

J. Bhagwati, V. K. Ramaswami, and T. N. Srinivasan, "Domestic Distortions, Tariffs, and the Theory of Optimum Subsidy: Some Further Results," *J. Polit. Econ.*, Nov. 1969, *77*, 1005–10.

J. Bhagwati and T. N. Srinivasan, "On Reanalyzing the Harris-Todaro Model: Policy Rankings in the Case of Sector-Specific Sticky Wages," *Amer. Econ. Rev.*, June 1974, *64*, 502–08.

G. Hadley and M. C. Kemp, *Variational Methods in Economics*, Amsterdam 1971.

J. Harris and M. Todaro, "Migration, Unemployment and Development: A Two-Sector Analysis," *Amer. Econ. Rev.*, Mar. 1970, *60*, 126–42.

H. G. Johnson, "Optimal Trade Intervention in the Presence of Domestic Distortions," in R. E. Caves et al., eds., *Trade, Growth, and the Balance of Payments*, Amsterdam 1965.

S. Magee, "Factor Market Distortions, Production and Trade: A Survey," *Oxford Econ. Pap.*, Mar. 1973, *25*, 1–43.

# Population Density, Agricultural Technique, and Land Utilization in a Village Economy

*By* J. DIRCK STRYKER*

Various writers have explored theoretically the range of economic decisions made by peasant families or communities in less developed countries. This has involved the linking of consumption and production choices (see Chihiro Nakajima and Don Winkelman) and the inclusion of nonagricultural productive activities (see Stephen Hymer and Stephen Resnick), land tenure relationships (see P. K. Bardhan and T. N. Srinivasan), and family planning (see Philip Neher). Attention has also focused on the risk inherent in agricultural activity or participation in the wage labor force (see Peter Gould, Michael Todaro, Jerome Wolgin).

An important dimension of peasant decisions which has not been fully explored is that of time. As Gary Becker has shown, the inclusion of this variable leads to important insights into family behavior. The common practice of dividing all time into two components, labor and leisure, however, inhibits the development of many of these insights. One exception has been consideration of the cost of travel time in studies of modern transportation systems, but there has been little attention to its cost in traditional economies.

Aside from the allocation of time, the other major variable over which the peasant family or community has control is the use of land. Most rural economy models assume that the amount of land available to each farmer is fixed and that he has only to decide the intensity of its use and the pattern of cropping.[1] Although this may be appropriate for densely populated regions, it is not applicable to areas of sparse population where land is in abundant supply. Due to the primitive means of transportation existent in such areas, the major constraint on the land brought into cultivation is not the boundaries of land belonging to others but rather the cost in terms of time of traveling between the fields and the farmer's place of residence.

This paper develops a model of an agricultural economy in which the area farmed is constrained only by the time required for this travel. With labor assumed to be the only other factor input, the intensity of land use is determined by the implicit cost of this travel in terms of time. Furthermore, the cost of using land depends not only on the area farmed by each family but also on the number of persons who choose to live together in a single hamlet or village. The model is applied both to subsistence production and to commercial agriculture, and the effects of changing population and relative prices are examined. Finally, the major hypotheses generated by the model are summarized and their implications assessed.

* Associate professor of international economic relations, Fletcher School of Law and Diplomacy, Tufts University. I would like to thank George Borts, Yoav Kislev, Harold Payson III, Stephen Resnick, and various members of the Economic Growth Center, Yale University, as well as an unknown referee, for their comments and suggestions.

[1] In some cases allowance is made for the existence of a land market, but it is usually assumed to be very imperfect. See, for example, Amartya Sen.

## I. A Subsistence Economy

It is assumed that the population of a given region is concentrated in nucleated villages from which farmers go to work in surrounding fields. Assuming no limit to land of homogeneous quality available for cultivation, its only cost is the time required by each farmer to go to and from his fields. If these fields are located so that the burden of travel time is shared equally by all farmers, the time required per farmer is given by a function $t(aN)$, where $a$ is the average amount of land cultivated by each and $N$ is the size of the village population, assumed to be the same as the work force. Since more travel time is required per farmer the larger is the area cultivated by each, the partial derivative of t with respect to $a$, $t_a$, is positive. The second-order partial derivative $t_{aa}$ may be either negative or positive.[2] The partial and cross-partial derivatives $t_N$ and $t_{aN}$ are reasonably assumed to be positive.

Assume further that the utility functions of all persons in the village are identical, that decisions are made by the village collectively, and that all individuals are treated in the same manner. All farmers have equal access to land, work the same number of hours, and employ the same technique.[3]

Per capita leisure $s$ is rather broadly defined to include all the time left over from a fixed number of hours $\bar{y}$ after subtracting agricultural labor $l$ and travel time t.[4] At first it is assumed that the village, with a fixed population size, seeks to maximize its leisure subject to the constraint that per capita agricultural output $x$ must equal the subsistence standard $\bar{x}$. The production function is given by $x(l, a)$ and we assume that $x_l$, $x_a$, $x_{al}>0$ and $x_{ll}$, $x_{aa}<0$, where subscripts denote first- and second-order partial derivatives.

Forming the Lagrangean expression $\bar{y}-l-t(a)-\lambda(\bar{x}-x(l, a))$ and setting the partial derivatives with respect to $l$, $a$, and $\lambda$ equal to zero, the following first-order conditions are obtained:[5]

$$-1 + \lambda x_l = 0 \tag{1}$$

$$-t_a + \lambda x_a = 0 \tag{2}$$

$$x(l, a) - \bar{x} = 0 \tag{3}$$

Equations (1) and (2) give the equilibrium condition

$$\frac{x_a}{x_l} = t_a \tag{4}$$

The ratio of the marginal products, or the marginal rate of technical substitution in production between labor and land, is equal to the marginal rate at which time can be exchanged for land by intravillage travel.

The equilibrium condition (4) is shown diagrammatically in Figure 1. For any

[2] The area of land expands more rapidly than the distance traveled as successive increments of land are brought into cultivation. If travel time is proportional to distance traveled, the cost in terms of travel time of bringing each new plot into cultivation tends to fall. Let $t=\gamma R$, where $\gamma$ is a constant and $R$ is the radius of a circle representing the total land area $A$ cultivated by the village. Since $A=\pi R^2$ and $a=A/N$, $t=\gamma'(Na)^{1/2}$, $t_a>0$ and $t_{aa}<0$. In order for the second derivative to be nonnegative, travel time must vary with at least the square of the distance traveled.

[3] This is consistent with an egalitarian type of tribal organization frequently found where land is abundant (see George Dalton).

[4] Hymer and Resnick have analyzed an agrarian economy in which the concept of leisure is extended to include such nonagricultural activities as food processing, housebuilding, recreation, and transportation. Since labor time is the only input assumed to be used in the production of these goods and services, and since we are not concerned with increases of labor productivity in this sector, we assume that welfare is a function of the inputs of time instead of the outputs resulting from its use outside of agriculture. This assumption also avoids some of the questions which Becker has raised concerning the effects of differences in the time intensity of consumption of various goods and services.

[5] Throughout the paper we are concerned only with solutions for which all variables take on nonnegative values.

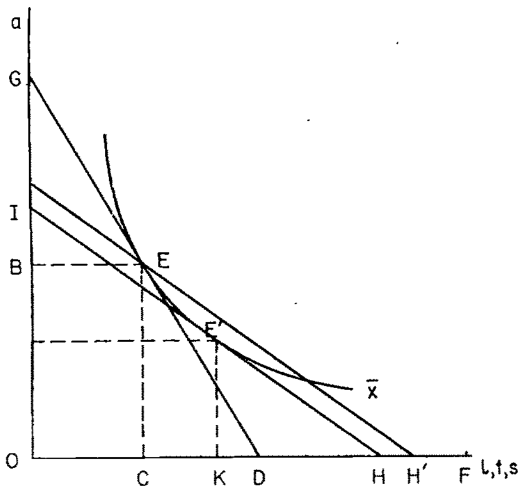



FIGURE 1

given value of $s$ and $N$ the equation $\bar{y}=s+l+t(a)$ is represented by a curve indicating the possible tradeoff between labor and agricultural land made available by the use of time for traveling. In Figure 1, $DG$ is such a curve corresponding to $DF$ of leisure, given the total time available $OF$.[6] The tangency at point $E$ of $DG$ and the isoquant corresponding to $\bar{x}$ represents the equilibrium condition of equation (4). In equilibrium $OC$ units of labor are combined with $OB$ units of land, requiring $CD$ units of travel time.

If equations (1)–(3) are differentiated totally, one obtains (5):

$$(5)\quad \begin{bmatrix} \lambda x_{ll} & \lambda x_{la} & x_l \\ \lambda x_{al} & -\mathrm{t}_{aa}+\lambda x_{aa} & x_a \\ x_l & x_a & 0 \end{bmatrix} \begin{bmatrix} dl \\ da \\ d\lambda \end{bmatrix} = \begin{bmatrix} 0 \\ \mathrm{t}_{aN}dN \\ 0 \end{bmatrix}$$

[6] For purposes of clarity and because of a lack of presumption of the direction of curvature, this equation is represented diagrammatically by a straight line. The analysis, however, does not depend on this assumption but only on satisfying second-order conditions discussed below.

In order for leisure to be maximized, it is a sufficient condition that the determinant $|D|$ of this system of equations be positive. This will occur as long as $\mathrm{t}_{aa}$, the curvature of $DG$, is either positive or, if it is negative, has an absolute value less than that of $d^2l/da^2$, the curvature of the isoquant at the point of equilibrium. Assuming that this second-order condition is satisfied, the effect of a change in population $N$ on $l$ and $a$ is seen by applying Cramer's rule:

$$(6)\qquad \frac{dl}{dN} = +\frac{\mathrm{t}_{aN}x_l x_a}{|D|} > 0$$

$$(7)\qquad \frac{da}{dN} = -\frac{\mathrm{t}_{aN}x_l^2}{|D|} < 0$$

An increase in $N$ thus changes the effective price of land in terms of time and causes the curve $DG$ in Figure 1 to be rotated about the given isoquant ($x=\bar{x}$) to $HI$, establishing a new equilibrium at $E'$.

From equations (6) and (7), $l$ increases and $a$ decreases as $N$ increases. What is the effect on the total area $A$, travel time t, and leisure time $s$? The value of $dA/dN$ is given by

$$(8)\qquad \frac{dA}{dN} = \frac{d(aN)}{dN} = N\frac{da}{dN} + a$$

Its sign will depend on the relative importance of the first term which is negative, and the second term which is positive. Similarly, the sign of $d\mathrm{t}/dN$ is given by

$$(9)\qquad \frac{d\mathrm{t}}{dN} = \mathrm{t}_N + \mathrm{t}_a\frac{da}{dN}$$

The first term is positive and indicates the increase in travel time from $CD$ to $CH'$ in Figure 1 as $N$ is increased with $a$ momentarily held constant. Intuitively, in order to accommodate the increased population each farmer must extend his fields outward if he is to farm the same total area of land. The second term shows the decrease in t

from $CH'$ to $KH$ as farmers decrease their use of land and increase that of labor along the isoquant $\bar{x}$ in response to the higher implicit price of using land.

The sign of $ds/dN$ is, however, always negative. To see this, solve equation (4) for $x_a$ and substitute both this and equation (7) in equation (6) to obtain

$$\frac{dl}{dN} = -\mathrm{t}_a \frac{da}{dN} \tag{10}$$

Differentiating the time constraint and solving for $ds/dN$,

$$\frac{ds}{dN} = -\frac{dl}{dN} - \frac{d\mathrm{t}}{dN} = -\mathrm{t}_N < 0 \tag{11}$$

since the second term of equation (9) is exactly offset by the value of $dl/dN$ from equation (10).

## II. A Commercial Economy

The transition from a purely subsistence economy to one in which most production is for commercial purposes is probably a gradual process. In the colonial empires it was often initiated by the imposition on the part of the colonial authorities of production quotas. Alternatively, a target income might be desired for the payment of taxes, a bride price, or the cost of some particular commodity. Ultimately, the expansion of demand to include a full range of goods and services originating outside the village implies the development of a utility function which includes not only leisure but also real income measured in terms of these goods and services.

It is possible to analyze each of these cases using extensions of the model. The enlargement of output due to the imposition of a quota is equivalent to an increase in $\bar{x}$. A target income might be acquired by selling surplus agricultural products, in which case the constraint becomes

$$\bar{m} = P(x(l, a) - \bar{x}) \tag{12}$$

where $\bar{m}$ is the real value of the target income and $P$ is the price ratio at which farm output may be exchanged for goods and services originating outside the village.

With the full development of commercial agriculture and of utility which is a function of real income as well as of leisure, the Lagrangian expression becomes $u(m, \bar{y}-l-\mathrm{t}(a))-\lambda(m-Px(l, a))$.[7] It is assumed that $u_m$, $u_s$, $u_{ms}>0$ and $u_{mm}$, $u_{ss}<0$. The first-order conditions now are

$$-u_s + \lambda P x_l = 0 \tag{13}$$

$$-\mathrm{t}_a u_s + \lambda P x_a = 0 \tag{14}$$

$$u_m - \lambda = 0 \tag{15}$$

$$Px(l, a) - m = 0 \tag{16}$$

yielding the equilibrium conditions

$$\frac{x_a}{x_l} = \mathrm{t}_a \tag{17}$$

$$\frac{u_s}{u_m} = Px_l \tag{18}$$

The first condition is the same as equation (4), where $x=\bar{x}$; the second condition states that the marginal rate of substitution between time and income in consumption must equal the marginal rate of substitution between time and income in production. These conditions are shown together diagrammatically in Figure 2. The upper part of that figure is similar to Figure 1. The lower part depicts a set of indifference curves with the origin at $F(\bar{y}=OF)$ and utility increasing in the southwest direction. For any function

[7] This form of the Lagrangian expression is used for convenience. Equivalent results are obtained if the expression takes the form $u(m,s)-\lambda_1(m-Px(l,a))-\lambda_2(s+l+\mathrm{t}(a)-\bar{y})$, but the analysis is mathematically more cumbersome. It is also implicitly assumed that all subsistence food requirements are met out of real income $m$. The results are substantially unchanged if, instead, the assumption is that only a surplus $(x-\bar{x})$ is sold after the village's own needs are satisfied. In this case, if $x<\bar{x}$ in equilibrium, we revert to the model described in the previous section.

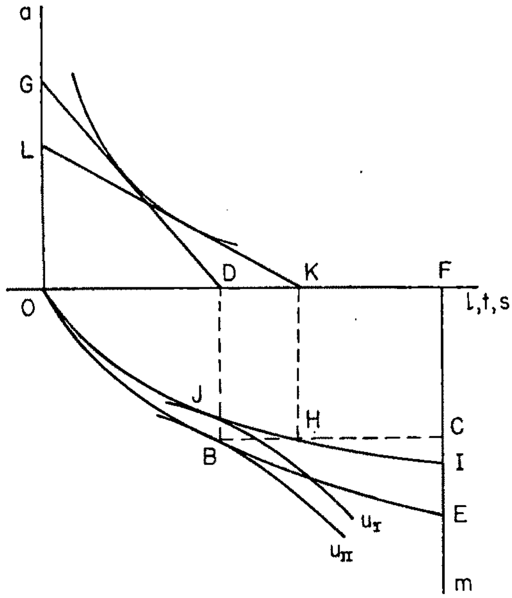



FIGURE 2

$t(aN)$ with $N$ given as a parameter of that function, there exists an equilibrium level of output for each amount of time spent in leisure. One such output multiplied by $P$ is denoted in the figure by point $B$, which corresponds to $DF$ units of leisure and $FC$ units of real income. Varying the amount of leisure traces out the curve $OE$. Differentiating $m=Px(l, a)$ and $s=\bar{y}-l-t(a)$,

$$\frac{dm}{ds} = -\frac{P(x_l dl + x_a da)}{d_l + t_a da} \tag{19}$$

Since in equilibrium $x_a = t_a x_l$, we can substitute for $x_a$ in the numerator to obtain an expression for the slope of $OE$:

$$-\frac{dm}{ds} = Px_l \tag{20}$$

Equation (18) is then represented by the tangency of $OE$ with the highest indifference curve, corresponding to a level of utility $u_{II}$.

If the area of land is fixed in supply so that $da=0$, equation (19) reduces directly to $-dm/ds=Px_l$ or $-dx/ds=x_l$. The difference between this situation and that represented by equation (20) is that in the former case $x_l$ is the slope of the usual marginal product curve of labor, while in the latter case the marginal product of labor falls less rapidly as $s$ is decreased and $l$ is increased because villagers are able to increase the total area farmed. This may be seen by examining the second derivative,

$$\frac{d^2x}{ds^2} = -\left(x_{ll}\frac{dl}{ds} + x_{la}\frac{da}{ds}\right) \tag{21}$$

If $da/ds=0$, $dl/ds=-1$ and $d^2x/ds^2=x_{ll}$. But if $da/ds<0$, the second term tends to offset the change in $x_l$ due to the first term. The value of $da/ds$ will depend, of course, on the function relating travel time to the area farmed and to the size of the village population as well as on the shape of the production function.

It is useful at this point to see what happens to the opportunity locus $OE$ when the village population changes. If $DG$ is rotated counterclockwise, indicating an increase in $N$, and kept tangent to the isoquant in Figure 2, leisure time is continually decreased for the given level of output as we move along $BC$. Point $H$ is determined, for example, by the curve $KL$. Any number of points could then be located for different levels of output and this new given level of $N$. These points all fall on the opportunity locus $OI$. The point of tangency between the indifference curve $u_I$ and $OI$ is the new equilibrium $J$.

We can examine the effects of $dN$ and $dP$ by differentiating equations (13)–(16) to obtain the system of equations shown as (22). One of the second-order conditions for maximization is satisfied if the determinant of this system is negative. We will

$$(22)\quad \begin{bmatrix} u_{ss}+\lambda P x_{ll} & t_a u_{ss}+\lambda P x_{la} & -u_{sm} & P x_l \\ t_a u_{ss}+\lambda P x_{al} & t_a^2 u_{ss}-t_{aa} u_s+\lambda P x_{aa} & -t_a u_{sm} & P x_a \\ -u_{ms} & -t_a u_{ms} & u_{mm} & -1 \\ P x_l & P x_a & -1 & 0 \end{bmatrix} \begin{bmatrix} dl \\ da \\ dm \\ d\lambda \end{bmatrix} = \begin{bmatrix} -t_N u_{ss} dN \\ t_{aN} u_s dN - t_a t_N u_{ss} dN \\ t_N u_{ms} dN \\ 0 \end{bmatrix} + \begin{bmatrix} -\lambda x_l dP \\ -\lambda x_a dP \\ 0 \\ -x dP \end{bmatrix}$$

assume this to be the case.[8]

The effects of a change in the village population are obtained by applying Cramer's Rule:

$$(23)\quad \frac{dl}{dN}=\frac{t_{aN}u_s D_{21}}{|D|} - \frac{t_N(u_{ss}D_{11}+t_a u_{ss}D_{21}-u_{ms}D_{31})}{|D|}$$

$$(24)\quad \frac{da}{dN}=\frac{t_{aN}u_s D_{22}}{|D|} - \frac{t_N(u_{ss}D_{12}+t_a u_{ss}D_{22}-u_{ms}D_{32})}{|D|}$$

$$(25)\quad \frac{dm}{dN}=\frac{t_{aN}u_s D_{23}}{|D|} - \frac{t_N(u_{ss}D_{13}+t_a u_{ss}D_{23}-u_{ms}D_{33})}{|D|}$$

where $D_{ij}$ is the cofactor of the element in the $i$th row and the $j$th column. The first term of each equation is the substitution effect resulting from the change in the implicit price of land in terms of time, initially holding per capita land, labor, and leisure constant. This term is negative for $da/dN$ and $dm/dN$.[9] It is usually positive for $dl/dN$ but could be negative if $x_{la}$ is very large. The second term shows the effect of an implicit change in the real value of total time resulting from a change in population and is always negative for $da/dN$. It is also negative for $dm/dN$ as long as $t_{aa}$ is either positive, or, if it is negative, has an absolute value less than that of $d^2l/da^2$, the curvature of the isoquant at the point of equilibrium. This was a second-order condition for maximization in the earlier model of a subsistence economy and is assumed to be satisfied here as well. Even if that condition is satisfied, however, the income effect for $dl/dN$ could still be positive if $t_{aa}$ is less than zero.

The effects of changing population on $A$, t, and $s$ may be shown by substituting equations (23) and (24) into (8), (9), and (11).

$$(26)\quad \frac{ds}{dN} = -\frac{dl}{dN}-\frac{dt}{dN} = -t_N - \frac{dl}{dN} - t_a\frac{da}{dN}$$

The terms of this equation are best understood by referring to Figure 3. Given per capita labor and land at the initial equilibrium point $M$, an increase in $N$ results in a decrease of leisure from $DF$ to $NF$, or a

[8] One term, $-u_m^2P^2(x_{ll}x_{aa}-x_{la}^2)$, in the determinant is positive if $x_{ll}x_{aa}<x_{la}^2$, a possibility if there are increasing returns to scale. The only other terms which could be positive are those including $Nt_{aa}$; these will be positive only if the value of the second derivative in these terms is negative. The second-order conditions might not be satisfied, then, in the event of strongly increasing returns to scale or a rapidly decreasing rate at which additional travel time is required as land area is increased.

[9] Explicit derivation of the signs of the terms described here is omitted, because of its mathematically cumbersome nature. Only the results are presented.

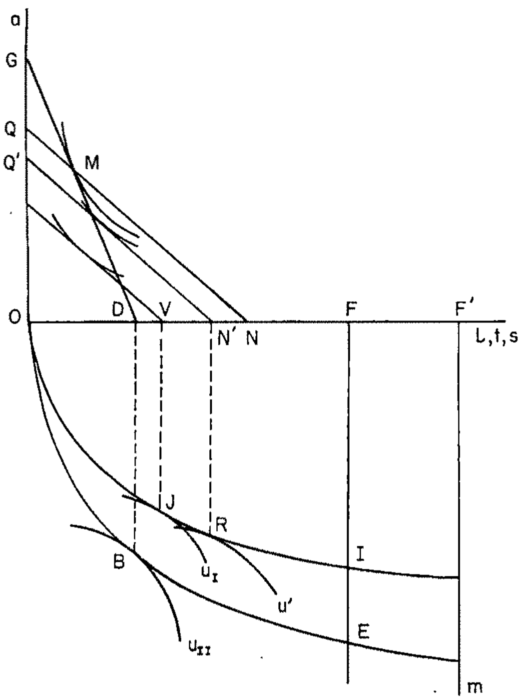



FIGURE 3

loss of $DN$. This decrease in leisure is represented by the first term $-t_N$ of equation (26). To show the substitution effect given in the last two terms of this equation, the total time available is increased by $FF'=DN$ to compensate for the loss of leisure due to the first term, and the indifference map is shifted to the right the same distance. The substitution effect, then, is shown by the movement from the initial equilibrium point $B$ to point $R$ where a shifted indifference curve is tangent to the new opportunity locus $OI$. Substituting equations (23) and (24) into equation (26), this substitution effect simplifies to

$$\frac{t_{aN}u_s u_m P}{|D|}(x_{ll}t_a - x_{la})$$

which is positive. Thus the initial leisure time represented by $DF=NF'$ is less than that indicated by $N'F'$ after the substitution effect has taken place. The movement along $OI$ from $R$ to $J$ is the income effect. Note that $J$ is the final equilibrium point representing a level of utility $u_I$ lower than that of $u_{II}$, as the indifference map is returned to its initial position. The income effect is negative for $ds$ as long as $t_{aa}$ is not negative and greater in absolute value than $d^2l/da^2$ along the isoquant, a possibility which has been ruled out. The final level of leisure represented by the distance $VF$ in Figure 3, for example, is less than $N'F'$. The total effect of a change in population on leisure will thus depend on the relative importance of the three terms of equation (26).

The effect of a change in $P$, perhaps due to an improvement in transportation to the village, is more straightforward. From the system of equations (22),

$$\text{(27)}\quad \frac{dl}{dP} = -\frac{u_m x_l D_{11} + u_m x_a D_{21}}{|D|} - \frac{xD_{41}}{|D|}$$

$$\text{(28)}\quad \frac{da}{dp} = -\frac{u_m x_l D_{12} + u_m x_a D_{22}}{|D|} - \frac{xD_{42}}{|D|}$$

Neglecting for the moment the influence of $t_{aa}$, the first term representing the substitution effect is positive in each case, and the income effect, or last term, is negative. Again, if $t_{aa}$ were negative, this would tend to offset or reverse the direction of these effects for $dl/dP$. The direction of each effect is unambiguous, however, for $dx/dP = x_l(dl/dP)+x_a(da/dP)$ as long as the condition stated earlier concerning the relationship between $t_{aa}$ and $d^2l/da^2$ continues to hold.

There is by now substantial evidence that where surplus land is available, peasant farmers have responded positively to price changes.[10] Assuming, then, that the substitution effect dominates the income effect, it is possible to compare the results shown by our model with the situation in

[10] See, for example, summaries of some of this evidence by William Jones and Raj Krishna.

which there is an absolute land constraint.

These two situations are depicted in Figure 4. The opportunity locus with land still available beyond the village is the curve $OE$. Alternatively, it may be assumed that the limits to village areal expansion are reached at point $B$, and, to the right of that point, output is given by the total product curve of labor with land held constant. This curve $BG$ lies everywhere inside of $BE$ from equation (21). To simplify the diagram, point $B$ has also been chosen as the initial equilibrium, but this assumption is not essential to the proof.

As a result of an increase in $P$, both $OE$ and $OG$ are shifted to $OC$ and $OH$, respectively. The indifference curves are drawn so that the substitution effect is greater than the income effect, and thus the new equilibrium points $I$ and $J$ lie to the right of point $B$.

If $$\left[\frac{dx}{dP}\right]_{da>0} \leq \left[\frac{dx}{dP}\right]_{da=0}$$

then $J$ must be located at, or to the right of, point $K$ along $OH$ since the level of output at $K$ is the same as at $I$. The slope of the indifference curve passing through $K$ is equal to $-u_s/u_m$. As we move from $K$ to $I$ the slope of indifference curves will decrease absolutely since $u_{ss}<0$ and $u_{ms}>0$.

But if $$\left[\frac{u_s}{u_m}\right]_K > \left[\frac{u_s}{u_m}\right]_I$$

neither $K$ nor any point along $OH$ to the right of $K$ can be a point of tangency since the slope of $OH$ at $K$ must be less than the slope of $OC$ at $I$.

Therefore, $$\left[\frac{dx}{dP}\right]_{da>0} > \left[\frac{dx}{dP}\right]_{da=0}$$

### III. Conclusions and Implications

The essential feature of the theoretical model developed in this paper is the constraint placed on the use of land, where it

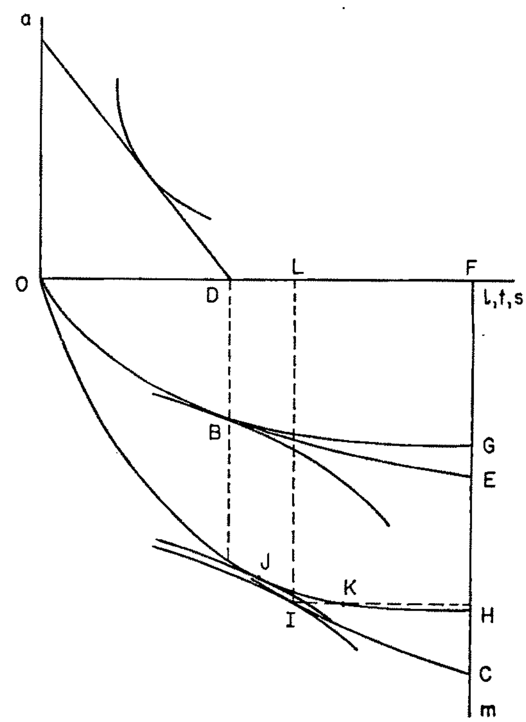



FIGURE 4

is otherwise freely available, by the time required to travel between cultivated fields and the place of residence. A major assumption is that people want to live together and are willing to sacrifice part of their scarce time in order to fulfill that desire. The degree of concentration of residences differs, however, between various societies. In some areas nuclear families live by themselves on individual farms; in others the villages or towns from which people travel to cultivate their fields are quite large.[11] The time spent traveling depends both on the means of available transport and on the relative importance which different societies attach to the frequency and extent of social contact.

Although evidence is very scattered,

[11] William Morgan has noted the existence in Africa, for example, of nucleated rural settlements of over 5000 people, requiring distances of five miles or more to be traveled regularly to reach the farthest fields.

that which exists indicates that this travel time is of considerable importance in many agricultural regions. Michael Chisholm, for example, finds that about one-third of total working hours are spent in traveling on the farm in regions as diverse as Western Europe and Nigeria. Detailed studies of rural areas by the Ministry of Planning in the Ivory Coast indicate that travel time in areas of sparse population density is between one-quarter and one-third of the total hours devoted to agriculture. These estimates are indicative of the orders of magnitude involved, but there is a need for much better information on variation in travel time between different regions and on how travel time is related to the area farmed. Because of the lack of empirical data necessary to specify the form of this relationship, it has been kept quite general in the preceding analysis.

Despite this high level of generality, however, there are a number of conclusions derived from the theory which can be summarized in the form of hypotheses susceptible to empirical verification and the implications of those hypotheses.

1. *The marginal product of land is positive and is equal to the marginal product of labor multiplied by the marginal rate at which travel time may be exchanged for land.*

Verification of this hypothesis would require estimation of both the production function and the function relating travel time to cultivated land. One implication of the hypothesis is that land is not a costless factor of production even where it is abundant, and that there may be gains from improving transportation methods within the village area—by the purchase of bicycles, for example. This improvement would result in a shift in the function involving t and *a* similar to that which would occur from a decrease in population, and thus in a substitution of land for labor. The net effect on output depends upon the type of economy, but there would be an improvement in welfare resulting from an increase in real income, leisure, or both.

2. *In a subsistence economy, an increase in population will lead to an increase in the amount of labor and a decrease in the amount of land used per person despite the availability of free land surrounding the village.*

The intensification of agriculture which accompanies a growth in population does not, therefore, depend on the limits to cultivable land. The degree of pressure exerted by population growth on land use depends on the degree of social cohesion acting to hold the village together. If the forces tending to concentrate living areas close together are strong and village transportation facilities are poor, means must be quickly employed to avoid soil depletion.[12] Where forces tending to concentrate living areas are less strong, however, relief from population pressure will occur as members of the village break away to establish new settlements outside the village perimeter.

This suggests that governments anxious to avoid soil deterioration due to population pressure in areas of otherwise abundant land should be aware of the size villages reach before significant numbers of farmers are willing to leave. It also suggests that the value of population concentration to farmers may be measured by the time required to travel to the farthest fields before they begin to found new villages, a factor which should be taken into account in the design of any resettlement schemes.[13]

[12] The problem is not so much the physical inability of farmers to maintain fertility of the soil despite shorter periods of fallow (see Ester Boserup), though this may be difficult in some ecological zones, but the speed with which new techniques must be adopted if population growth is very rapid.

[13] Chisholm, for example, finds a surprisingly consis-

3. *In a cash crop economy, an increase in population will result in a decrease in the amount of land used per capita and a decrease in the amount of per capita output despite the availability of free land surrounding the village.*

Since both income and substitution effects operate in the same direction, this conclusion is unambiguous. It implies that population growth will cause peasant farmers to shift away from production for cash, and, if carried far enough, could cause them to return to subsistence farming. The point at which villagers break off to form new settlements is of course an important determinant of the effect of population growth on production for market.

4. *In a subsistence economy, the effect of population growth is to reduce the time allocated to leisure or to nonagricultural productive activities. In a cash crop economy, the net effect on this time depends on the relative importance of terms with differing signs. There is, however, an unambiguous decrease in welfare.*

The adverse effects on welfare of population growth and increasing pressure on the land are not, therefore, confined to densely populated countries.

5. *Assuming that there is a positive production response to a change in the relative price of the cash crop, that response is greater where there is free land available than where land is fixed in supply.*

There is plentiful evidence that very rapid rates of growth in cash crop production have occurred historically in some areas of abundant land in response to the stimulus of improved transportation or favorable marketing situations.[14] Although immigrant labor has frequently been responsible for the continuation of this expansion, it has usually been initiated by the local peasantry. This growth has occurred, moreover, as a result of increases both in labor hours and in the area of land cultivated by each farmer.[15]

Although the introduction of commercial agriculture into these countries has been notably successful, there are other areas with abundant land resources where no such success has been achieved.[16] While there may be many reasons for this, such as lack of infrastructure and local agricultural experimentation as well as restrictive government policies of various sorts, my analysis suggests that attitudes towards leisure, the size of villages, the strength of forces promoting village concentration, and the degree of difficulty of traveling between village and fields may have been among the contributing factors. Where the expansion of commercial agriculture by local farmers for these reasons is unlikely, it may be possible to encourage cultivation by immigrants, provided institutional arrangements can be made for these people to obtain access to land.[17]

tent 3 to 4 kilometers of maximum travel distance to be the threshold of resettlement in many parts of the world, though there are instances in which a new settlement is established after only $1\frac{1}{2}$ or 2 kilometers is reached. Morgan, on the other hand, believes the limiting distance in Africa is 4 to 7 miles.

[14] Among the countries which have achieved these high rates of "vent-for-surplus" growth (see Hla Myint) during certain periods in the past are Burma, Ghana, Malaysia, Nigeria, South Vietnam, Thailand, and Uganda. Probably the most notable recent example is the Ivory Coast, which has sustained an average annual rate of growth of exports of agricultural and forest products since World War II of about 8 percent.

[15] See, for example, Ivory Coast, pp. 70–76, and Godwin Okurume, p. 314.

[16] Along the relatively sparsely settled coast of the Gulf of Guinea in West Africa, for example, Nigeria, Ghana, and the Ivory Coast have experienced rapid rates of growth of peasant cash crop production during various periods of their history whereas Dahomey, Togo, Liberia, Guinea, and Sierra Leone have not.

[17] The classic case of this in Africa occurred in Southern Ghana, as documented by Polly Hill.

## IV. Limitations and Extensions

The theoretical model presented here is limited to forms of traditional agriculture in which land and labor are the only inputs. Changes in agricultural technique are due to movements along a production function which is already known, at least in the neighborhood of existing methods. There is no allowance for technological improvement, which would be represented here by a shift of the production function,[18] or for the introduction of modern industrial inputs, such as chemical fertilizers. Traditional forms of capital have also not been included as an input. These could be considered as embodied forms of land and labor, leaving the conclusions of the analysis unaffected in the long run. This prevents us, however, from explicitly considering aspects of capitalization which might be related to distance within the village. In addition, it has been assumed that there is one homogeneous crop and that a single technique is employed at any given time. To conserve labor and travel time, however, several techniques might be employed simultaneously, either for the same product or for a variety of crops. Finally, the crucial assumption is made in the model that land is of homogeneous quality. To the extent that there are variations in quality, whether related or not to distance, this will obviously affect some of the results.

Extensions of the model might be directed toward the relaxation of these assumptions. Of particular importance would be the introduction of off-farm inputs, so crucial to agricultural development. There is evidence, for example, that the value of output per acre net of input costs declines much more rapidly with distance from the village center than does that of gross output.[19] Different inputs are likely to differ, furthermore, in the travel time intensity of their use. Where several crops can be cultivated using a variety of techniques, it is likely that more labor intensive methods will be employed closer to the village.[20] It is also likely that villages are situated in such a way that soil fertility declines as distance from the village center increases. Variations in the quality of land should, consequently, be introduced.

In addition, travel time is related to factors other than land area. The frequency of trips to different plots of land is often associated, for example, with seasonal factors such as the need to insure protection of crops from wild animals or birds during crucial stages of growth. Food crops, such as vegetables and herbs which may be harvested at intermittent intervals, also require more frequent trips than other crops such as grains, which are harvested within a single, fairly short space of time. Application of the model to individual situations should take into account such peculiarities.

This paper has shown that although there are advantages associated with free land, they are not costless. Unexplained, however, are many of the disadvantages related to low population density, such as the high cost of providing transportation, commercial, and government services to people in these areas. The immediate advantages of sparsely populated regions may, in fact, be outweighed by the poorer prospects for commercial agriculture, wage employment, education, and health where densities are low.

[18] Or, in terms of the analysis of Yujiro Hayami and Vernon Ruttan (pp. 82–84), a movement along a "metaproduction function," or "potential production function," which describes the technical alternatives which could be discovered given the existing fund of general scientific knowledge.

[19] See Chisholm, pp. 49–52.

[20] This pattern of cultivation has been widely observed throughout the world by such writers as W. Allen, Chisholm, Morgan, Karl Pelzer, and M. B. Gleave and H. P. White.

## REFERENCES


W. Allan, *The African Husbandman*, Edinburgh 1965.

P. K. Bardhan and T. N. Srinivasan, "Crop-sharing Tenancy in Agriculture: A Theoretical and Empirical Analysis," *Amer. Econ. Rev.*, Mar. 1971, *61*, 48–64.

G. S. Becker, "A Theory of the Allocation of Time," *Econ. J.*, Sept. 1965, *75*, 493–517.

E. Boserup, *The Conditions of Agricultural Growth; The Economics of Agrarian Change Under Population Pressure*, Chicago 1965.

M. Chisholm, *Rural Settlement and Land Use*, Chicago 1970.

G. Dalton, "Traditional Production in Primitive African Economies," *Quart. J. Econ.*, Aug. 1962, *76*, 360–78.

M. B. Gleave and H. P. White, "Population Density and Agricultural Systems in West Africa," in M. F. Thomas and G. W. Whittington, eds., *Environment and Land Use in Africa*, London 1969, 273–300.

P. R. Gould, "Man Against his Environment: A Game Theoretic Framework," *Annals Assn. American Geographers*, Sept. 1963, *53*, 290–97.

Y. Hayami and V. W. Ruttan, *Agricultural Development: An International Perspective*, Baltimore 1971.

P. Hill, *The Migrant Cocoa-Farmers of Southern Ghana; A Study in Rural Capitalism*, Cambridge 1963.

S. Hymer and S. Resnick, "A Model of an Agrarian Economy with Nonagricultural Activities," *Amer. Econ. Rev.*, Sept. 1969, *59*, 493–506.

W. O. Jones, "Economic Man in Africa," *Food Res. Inst. Stud.*, May 1960, *1*, 107–34.

R. Krishna, "Agricultural Price Policy and Economic Development," in H. M. Southworth and B. F. Johnston, eds., *Agricultural Development and Economic Growth*, Ithaca 1967, 497–540.

W. B. Morgan, "The Zoning of Land Use Around Rural Settlements in Tropical Africa," in M. F. Thomas and G. W. Whittington, eds., *Environment and Land Use in Africa*, London 1969, 301–19.

H. Myint, "The 'Classical Theory' of International Trade and the Underdeveloped Countries," *Econ. J.*, June 1958, *68*, 317–37.

C. Nakajima, "Subsistence and Commercial Family Farms: Some Theoretical Models of Subjective Equilibrium," in C. R. Wharton, Jr., ed., *Subsistence Agriculture and Economic Development*, Chicago 1969, 165–85.

P. A. Neher, "Peasants, Procreation, and Pensions," *Amer. Econ. Rev.*, June 1971, *61*, 380–89.

G. E. Okurume, "Foreign Trade and the Subsistence Sector in a Peasant Economy; A Case Study of the Impact of Agricultural Export Expansion on Domestic Food Supplies in Western Nigeria," unpublished doctoral dissertation, Yale Univ. 1970.

K. J. Pelzer, *Pioneer Settlements of the Asiatic Tropics; Studies in Land Utilization and Agricultural Colonization in Southeastern Asia*, New York 1948.

A. K. Sen, "Peasants and Dualism With or Without Surplus Labor," *J. Polit. Econ.*, Oct. 1966, *74*, 425–50.

M. P. Todaro, "A Model of Labor Migration and Urban Unemployment in Less Developed Countries," *Amer. Econ. Rev.*, Mar. 1969, *59*, 138–48.

D. Winkelman, *The Traditional Farmer: Maximization and Mechanization*, Paris 1972.

J. M. Wolgin, "Response to Price in Smallholder Agriculture in Kenya: A Cross-Sectional, Portfolio-Choice Model," unpublished doctoral dissertation, Yale Univ. 1973.

Ivory Coast, Ministère du Plan, *Région du Sud-Est; Etude Socio-Economique: L'Agriculture I*, Paris 1967.

# Income Transfers as a Public Good: An Application to AFDC

By LARRY L. ORR*

One of the dominant features of the U.S. income transfer system is the great disparity in benefits available to similarly situated persons in different political jurisdictions. Public assistance (welfare) benefits in the most generous states are over six times as large as those in the least generous states. In recent years, this observation has increasingly provided the impetus for a movement to federalize the entire welfare program, and to establish a uniform federal benefit schedule in all states.

Unfortunately, the policy debate over federalization of welfare has proceeded in the context of an almost total ignorance of what—other than "liberality"—determines the level of income transfers in the various states under the current system.

Only recently have economists begun to develop a consistent rationale for the existence of income transfers,[1] and none of the existing theoretical models has been tested empirically. Yet, even under a federalized system, some disparities are likely to remain, since states would be free to supplement the federal payments (albeit without federal assistance), and some presumably would. Thus, it is important to understand the factors which give rise to disparate transfer levels, whether or not a wholly federal program is envisioned.

This study attempts to provide a theoretically consistent model of the federal-state transfer system in the context of a democratic decision-making framework, and provides an empirical test of that model for the largest of the existing welfare categories, Aid to Families with Dependent Children (*AFDC*). These results are useful both for understanding the structure of benefits in the existing program, and for predicting the effects on benefit levels of a variety of welfare policy options.

## I. A Theoretical Model of Redistribution

The theoretical framework treats income redistribution as a public good. The assumed rationale for redistribution is that taxpayers derive utility from increments to the income of transfer recipients. Since each taxpayer receives utility from all transfers made, but is required to pay for only a portion of the aggregate transfer cost, the theory of public goods can be applied (with certain modifications) to determine the equilibrium level of transfers. While differing in important respects from earlier models, the discussion here is very much in the spirit of the transfer models developed by Pauly, Olsen, Goldfarb, and von Furstenberg and Mueller.

In order to keep the analysis manageable, several rather stringent assumptions are made. It is assumed that:

* Director, Office of Income Security Policy Research, U.S. Department of Health, Education, and Welfare. The research reported herein benefited greatly from comments and suggestions by Edward Gramlich, Adele Massell, Irwin Garfinkel, and anonymous referees. I am also indebted to Mike Hallstrom and Jane Lee for computational assistance. I bear sole responsibility for the analysis, views, and conclusions expressed herein. Nothing contained herein represents the official policy of the U.S. Department of Health, Education, and Welfare. 

[1] See, e.g., George von Furstenberg and Dennis Mueller; Robert Goldfarb; Edgar Olsen; Mark Pauly; Harold Hochman and James Rogers; Lester Thurow; and Gordon Tullock.

1) There are two mutually exclusive groups within the transfer jurisdiction, taxpayers and recipients.

2) Transfers are financed from taxes, allocated (at the margin) among taxpayers according to exogenously determined tax shares. For simplicity, it will be assumed that all taxpayers pay an equal share of the cost of transfers.

3) All recipients receive equal amounts of transfer income.

4) Recipients' incomes enter the utility function of each taxpayer; the incomes of taxpayers enter neither the utility functions of recipients nor of other taxpayers. The first partial derivatives of each taxpayer's utility with respect to both own-income and recipients' incomes are all positive; the second partials are all negative.

5) The level of transfers is determined by majority vote of the residents of the transfer jurisdiction, including recipients.

We begin by deriving the level of transfers which would be deemed optimal by a single taxpayer with given income and tastes. By assumption 3), we can characterize the level of transfers by a single variable, $B$, the (uniform) transfer benefit per recipient per time period. By assumption 4), the $i$th taxpayer has a utility function $U_i$ of the form:

$$U_i = U_i(Y_i, Y_j) \tag{1}$$

where $Y_i$ is the after-tax income of the $i$th taxpayer, and $Y_j$ is the vector of after-transfer incomes of recipients. The utility gain (loss) to this taxpayer from a unit increase in the per recipient level of transfer $B$, is:

$$\frac{\partial U_i}{\partial B} = \frac{\partial U_i}{\partial Y_i}\cdot\frac{\partial Y_i}{\partial T_i}\cdot\frac{\partial T_i}{\partial B} + \sum_{j=1}^{p} \frac{\partial U_i}{\partial Y_j}\cdot\frac{\partial Y_j}{\partial B} \tag{2}$$

where $T_i$ is the tax liability of the $i$th taxpayer. By the assumption of equal tax shares:

$$\frac{\partial T_i}{\partial B} = \frac{p}{n} \tag{3}$$

if there are $n$ taxpayers and $p$ recipients. Setting $\partial Y_j/\partial B = 1$, $\partial Y_i/\partial T_i = -1$, and substituting $p/n$ for $\partial T_i/\partial B$, equation (2) becomes

$$\frac{\partial U_i}{\partial B} = \sum_{j=1}^{p} \frac{\partial U_i}{\partial Y_j} - \frac{\partial U_i}{\partial Y_i}\cdot\frac{p}{n} \tag{4}$$

The first-order condition for a utility maximum for the $i$th taxpayer is obtained by setting equation (4) equal to zero, or:

$$\sum_{j=1}^{p} \frac{\partial U_i}{\partial Y_j} = \frac{\partial U_i}{\partial Y_i}\cdot\frac{p}{n} \tag{5}$$

Figure 1 presents this optimality condition graphically. The function *mb* (marginal benefit) in Figure 1 gives the value of the left-hand term of equation (5) for each level of $B$. It is monotonically decreasing because by assumption (4), the taxpayer's marginal utility of transfers to each of the $p$ recipients is monotonically decreasing. The schedule *mc* (marginal cost) shows the value of the right-hand term of equation (5) as a function of $B$. It is monotonically increasing because the taxpayer has decreasing marginal utility of own-income and, therefore, his marginal utility of own-

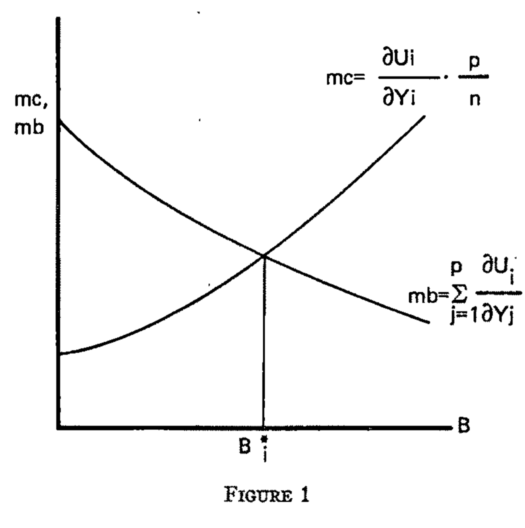



FIGURE 1

income rises with $B$ and his tax liability $T_i$.

The level of transfers which maximizes the taxpayer's utility, given his tastes, pretax income, and tax share, is given by the intersection of the two curves at $B_i^*$. At values of $B$ less than $B_i^*$, an increase in $B$ will raise his utility. At values of $B$ greater than $B_i^*$, a decrease in $B$ will increase his utility.[2]

If tastes and incomes vary among taxpayers, each taxpayer will in principle prefer a different level of transfers. Yet only one level of $B$ can be selected as public policy. This is simply the public good problem. The traditional theoretical approach to this problem has been to attempt to vary tax shares in such a way that there is unanimous agreement upon public transfer policy. (See, for example, Thurow.) Here we will adopt the less elegant (but probably more realistic) approach of assuming that differences of opinion are not eliminated by the political process; rather, a single value of $B$ is chosen as public policy by simple majority rule.

Thus, the utility maximization of individual taxpayers generates a distribution of preferred transfer levels given by the vector $B_i^*$. In addition, since recipients can vote even though they pay no taxes, the political process must take account of their preferences as well. Since increases in $B$ entail no costs but confer positive benefits for recipients, we may set their preferred level of transfers at infinity. The inputs to the political process, then, are the $(n+p)$ elements of the augmented vector $B_i^*$, where $p$ of these elements are infinite. Majority rule will select as public policy a value of $B$ equal to the median of this distribution. If any higher transfer level is chosen, a majority can be formed to lower $B$; similarly, if a lower level is set, a majority (including the recipients themselves) will exist to raise it.[3] The median of $B_i^*$, which we will denote $B^*$, is therefore a stable equilibrium level of transfers under majority rule, even though it is likely that many—indeed most—citizens would prefer a higher or lower level of redistribution.

So far, the discussion has implicitly assumed that the entire cost of transfers is borne by the taxpayers within the political jurisdiction where the transfers are made. The model can easily be extended, however, to allow for intergovernmental matching grants, whereby voters in one jurisdiction can in effect tax residents of other jurisdictions. Suppose, for example, that the federal government contributes to the cost of transfers (per recipient) in each state according to a fixed matching schedule $F(B)$, which is a function of the benefit level in the state. Assuming equal tax shares at both state and federal levels, the taxpayer's total tax liability, for transfers in all states, will be:

$$T_i = \frac{(B - F(B))p}{n} + \frac{F(B)p}{N} + \frac{F(B^0)p^0}{N} \tag{6}$$

where $p^0$ and $B^0$ are, respectively, the total number of recipients and the average transfer level in all other states, and $N$ is the total number of taxpayers in all states. Differentiating this equation to obtain the value of $\partial T_i/\partial B$ in equation (2), we have:[4]

[2] Note that while the analysis here assumes that the taxpayer's tax share is fixed exogenously, the assumption of equal tax shares is not essential to the analysis. Any fixed tax share $s_i$ could be substituted for $p/n$ in equation (5) and Figure 1. In general, however, tax shares may be expected to vary positively with $p/n$.

[3] A further necessary assumption if the median level of $B$ is to be finite is that recipients always comprise a minority of the electorate. It is also necessary to assume that there are no "side payments"; i.e., votes are not bought or sold. Otherwise, the equilibrium solution may be indeterminate. Use of the median as a decision criterion in collective choices is analyzed in Duncan Black; application of this criterion in similar contexts has been suggested by Gary Becker, p. 63, and by David Bradford and Wallace Oates, pp. 443–44.

[4] This relationship implicitly assumes that the level of transfers in other jurisdictions is unaffected by $B$.

$$(7)\qquad \frac{\partial T_i}{\partial B} = \frac{(1 - F'(B))p}{n} + \frac{F'(B)p}{N}$$

where $$F'(B) = \frac{\partial F(B)}{\partial B}$$

Substituting this expression for $\partial T_i/\partial B$ in equations (2)–(5) yields:

$$(8)\qquad \sum_{j=1}^{p} \frac{\partial U_i}{\partial Y_j} = \frac{\partial U_i}{\partial Y_i} \cdot \left[\frac{(1 - F'(B))p}{n} + \frac{F'(B)p}{N}\right]$$

as the optimality condition for the $i$th taxpayer in the presence of federal matching.

In terms of the graphical representation of the optimality condition in Figure 1, the *mb* curve which depends only on the taxpayer's utility function is unaffected by federal matching. The *mc* curve, however, is shifted by both a price and an income effect when matching is introduced. The price effect of matching shifts the *mc* curve downward, as compared with the no-matching case. If $N$ is very large relative to $n$, so that the term $F'(B)/N$ is negligibly small, the level of the *mc* curve is approximately proportional to the term $(1-F'(B))$. This term is just the marginal state share of benefits; it may be viewed as the marginal price of benefits (per recipient) to taxpayers within the state. (Empirically, in the *U.S.* welfare system, the marginal state share is an increasing function of $B$, reinforcing the positive slope of *mc*.) Increases in the marginal federal matching rate, then, will lower the *mc* curve and increase the equilibrium level of $B_i^*$.

Matching will also have an effect on *mc* through its effect on the taxpayer's after-tax income $Y_i$. Consider, for example, a change in the matching formula which leaves the marginal matching rate unchanged but raises (lowers) the average federal share of benefits. Without loss of generality, we may represent the matching formula by:

$$(9)\qquad F(B) = a + f(B)$$

The effect of a parametric shift in the intercept of this function on the marginal utility of own-income is:[5]

$$(10)\qquad \frac{\partial\left(\dfrac{\partial U_i}{\partial Y_i}\right)}{\partial a} = \frac{\partial^2 U_i}{\partial Y_i^2} \cdot \frac{\partial Y_i}{\partial a} = \frac{\partial^2 U_i}{\partial Y_i^2} \cdot \left[-\frac{\partial T_i}{\partial a}\right] = \frac{\partial^2 U_i}{\partial Y_i^2} \cdot \left[\frac{p}{n} - \frac{p}{N}\right]$$

Under the assumption of decreasing marginal utility of own-income, this expression must be negative. Therefore, the income effect of an upward shift in $F(B)$ will lower the *mc* curve, *ceteris paribus*, and a downward shift in $F(B)$ will raise the *mc* curve. The resulting changes in the equilibrium transfer level, $B_i^*$, will be positive and negative, respectively.

Finally, it should be noted that an individual taxpayer may not value transfers to all recipients equally. In general, $\partial U_i/\partial Y_j$ need not equal $\partial U_i/\partial Y_k$ for any two recipients $j$ and $k$, although by assumption they must receive equal transfers. It may well be that taxpayers systematically favor (i.e., $\partial U_i/\partial Y_j$ may be higher for) certain classes of recipients as opposed to others. To the extent that the recipient population is comprised of these

[5] Equation (10), obtained by partial differentiation of equation (6), holds constant the federal matching rate in all other states. Implicitly, then, we are evaluating only the income effect of matching rate *differentials* among states. This is consistent with the empirical tests of the following section. Clearly, a change in the matching rate applicable to a single state will have a larger income effect on transfer levels in that state than would a simultaneous change in the matching rates in all states.

favored groups, the *mb* schedule in Figure 1, and therefore $B_i^*$, will be higher than it would otherwise be.[6]

We can now summarize the determinants of changes in the level of transfers over time and/or variations in transfer levels among states. As between two different states, or the same state at two different points in time, holding everything else constant:

1. Higher income among taxpayers will lower the *mc* curve (because of diminishing marginal utility of own-income) and result in higher transfers.
2. A lower ratio of recipients to taxpayers will lower the *mc* curve and result in higher transfers.
3. A larger absolute number of recipients (holding constant the recipient/taxpayer ratio) will raise the *mb* curve and result in higher transfers, since $\partial U_i/\partial Y_j$ is assumed to be positive, and *mb* is the summation of $\partial U_i/\partial Y_j$ over all recipients.
4. A higher federal matching rate will lower the *mc* curve (via both a price and an income effect) and result in higher transfers.
5. More altruistic tastes on the part of taxpayers (holding taxpayer income and the composition of the recipient population constant) in the form of higher $\partial U_i/\partial Y_j$ will raise the *mb* curve and result in higher transfers.
6. A recipient population with a larger proportion of individuals for whom taxpayers have a high $\partial U_i/\partial Y_j$ will raise the *mb* curve and result in higher transfers.

While the analysis from which these results are derived is carried out at the individual taxpayer level, it should be noted that under the political process of majority rule assumed here they also carry over to the aggregate public policy outcome. That is, a given change in any of these variables for all taxpayers changes all of the elements of $B_i^*$ in the same direction and, therefore, changes the median value of $B_i^*$ in the same direction.

## II. Empirical Tests of the Model for the *AFDC* Program

The largest *U.S.* transfer program designed to serve a purely redistribution function is Aid to Families with Dependent Children (*AFDC*). In January 1974, there were 10.4 million *AFDC* recipients.[7] The levels of assistance payments under *AFDC* are set by the various states which administer the program. The cost of *AFDC* payments is borne jointly by the federal and state governments under two different matching formulas. Under the original public assistance matching formula, the federal government agreed to pay 83 percent of the first $18 per recipient per month, 50–65 percent (depending on state per capita income) of the next $14, and nothing for that portion of benefits in excess of $32 per recipient per month. Since 1965, however, states have been allowed the option of straight proportional federal sharing at the matching rate established for the Medicaid program. This rate varies

[6] Where there are strong and widely shared differences in taxpayers' perceptions of the "deservingness" of different groups, and it is possible to discriminate among recipients on a socially acceptable basis, the equal-benefit constraint assumed here can be circumvented by categorizing recipients into separate programs with different transfer levels. Thus, we have separate transfer programs for the aged, the incapacitated, the unemployed, Cuban refugees, disabled coal miners, veterans, single-parent families with children, and two-parent families with children. In other cases, discriminatory taxpayer preferences may be just as widely shared, but either the basis for discrimination is not a socially or legally acceptable one (e.g., race), or discrimination is administratively infeasible (e.g., as between the "voluntary" and the "involuntary" poor). Even in these cases where both groups are entitled to the same transfers, however, the common transfer level will be affected by taxpayers' perception of the composition of the recipient population.

[7] This figure for *AFDC* recipients, and the variables throughout this section, excludes recipients of assistance in the "unemployed parent" segment of *AFDC*. In January 1974, there were 439,000 such recipients.

TABLE 1—INDEPENDENT REGRESSION VARIABLES

| Variable | Definition | Proxy for: |
|---|---|---|
| *Y* | State per capita income in 1967 dollars | Real taxpayer income |
| *RPOP* | Ratio of total *AFDC* recipients to state civilian population, lagged one year | Recipient/taxpayer ratio |
| *RCPT* | Total *AFDC* recipients in 1000's, lagged one year | Absolute number of recipients |
| *MP* | Marginal state share of *AFDC* payments, as a fraction | Federal matching rate (marginal price effect) |
| *F* | Federal share of average annual *AFDC* payment to family of four, in 1967 dollars | Federal matching rate (income effect) |
| *NE, OS, BS, W* | Dummy variables for Northeast, Old South, Border States, and West | Taxpayer preferences and other regional effects |
| *NWH* | Fraction of *AFDC* families headed by nonwhites | Composition of recipient population |
| T | Time, in calendar years (1962 = 1) | Time trend (linear) |
| TSQ | $T^2$ | Time trend (non-linear) |

inversely with state per capita income, with a minimum of 50 percent and a maximum of 83 percent. As of January 1974, all but twelve states had elected to adopt this matching formula rather than the original *AFDC* formula.

The *AFDC* program provides an excellent test of the theory of redistributive transfers developed in the previous section.[8] It is a large, well-established program with an open-ended federal matching commitment of the type embodied in the theoretical model. Moreover, virtually complete state autonomy in the determination of benefits has resulted in a wide range of transfer levels, reflecting, presumably, the variation in the underlying determinants of assistance levels among the states.

[8] Several previous studies have employed cross-sectional regression analysis to analyze the determinants of state public assistance expenditures. See, e.g., Lora Collins; Peter Albin and Bruno Stein; Solomon Fabricant; Glenn Fisher; Seymour Sacks and Robert Harris; Jack Osman; and Ronald Gold. These studies have not, however, been based upon any explicit theoretical model of the transfer decision-making process, and none has dealt in a satisfactory manner with the impact of federal matching grants. Moreover, only Collins has dealt with the various public assistance programs separately, and most of the other studies have sought to explain welfare expenditures per capita, rather than the level of transfers per recipient.

The preceding section advanced six major variables as theoretical determinants of the level of transfers in a redistributive program: taxpayer income, the recipient/taxpayer ratio, the absolute number of recipients, the federal matching rate, taxpayers' preferences, and the composition of the recipient population. The analysis reported below employs multiple regression analysis to estimate state *AFDC* payment levels as a linear function of these variables or their empirical proxies. The data base includes annual observations for all states and the District of Columbia for the period 1963–72.

The basic dependent variable selected for analysis is the average real (1967 dollars) monthly payment per *AFDC* recipient. This variable was scaled upward by a factor of 48, to represent the annual payment to a family of four. This transformation was performed merely to facilitate discussion of the results in the usual terms of annual transfers to the modal family. The transformed variable is denoted *B* in the analysis.

The independent variables employed in the analysis and their relationship to the variables of the theoretical model are displayed in Table 1. Real state per capita

income ($Y$) is entered as a measure of taxpayer income. Lagged values of the recipient/population ratio ($RPOP$) and the absolute number of recipients ($RCPT$) are used, in recognition of the likelihood that it takes states some time to adjust to changes in these variables. Two separate matching variables ($F$ and $MP$) are included to capture the income and marginal price effects of matching on benefit levels. Regional dummies ($NE$, $OS$, $BS$, and $W$) are included as proxies for taxpayer preferences with respect to redistribution (and any other regional effects). The racial composition of the caseload ($NWH$) was selected as the most important characteristic upon which taxpayers might discriminate among recipients.[9] Linear and quadratic time terms are included to capture any time trend not accounted for by the other variables.

The relationship we wish to estimate (with the expected signs of the explanatory variables) is:

$$(11)\quad \begin{aligned} B = {} & a_0 + a_1Y - a_2RPOP + a_3RCPT \\ & - a_4MP + a_5F \pm a_6NE \pm a_7OS \\ & \pm a_8BS \pm a_9W - a_{10}NWH \\ & \pm a_{11}T \pm a_{12}TSQ + e \end{aligned}$$

Unfortunately, direct estimation of this equation by ordinary least squares would yield biased estimates of $a_4$ and $a_5$, the coefficients which measure the price and income effects of matching. As noted earlier, under the original public assistance matching formula the marginal state share of *AFDC* payments ($MP$) rises as $B$ increases. The independent variable $F$ is also related to the dependent variable by the equation:

$$(12)\quad F = mB$$

where $m$ is the average federal matching rate. Both variables will therefore be positively correlated with the error term in equation (11), and the ordinary least squares estimates of $a_4$ and $a_5$ will be biased upward.[10]

In order to obtain unbiased estimates of $a_4$ and $a_5$, a two-stage estimation procedure was employed here. An estimated value of $F$ was obtained from the regression:

$$(13)\quad \begin{aligned} F &= \sum_i mX_ib_i + u \\ &= \hat{F} + u \end{aligned}$$

where the $X_i$ are the independent variables (other than $F$) of equation (11), $m$ is the average federal matching rate, and the $b_i$ are least squares regression coefficients. The variable $F$ was then replaced with $\hat{F}$ in estimating equation (11).[11] A similar

[9] Data on the racial composition of state *AFDC* caseloads were available only for the years 1961, 1967, 1969, and 1971, in which the National *AFDC* Surveys were conducted. Values of $NWH$ for other years were obtained by linear interpolation or (for 1972) extrapolation. In the last two surveys, state level data on racial composition are available only for 12 and 20 states, respectively. For the remaining states, the 1967 values were extrapolated on the basis of changes in the corresponding census division. This variable is, therefore, subject to an unknown degree of measurement error, which may bias its regression coefficients toward zero.

[10] Estimation of equation (11) by ordinary least squares for the periods 1963–67 and 1968–72 yields values of 1.04 and 1.55 for $a_5$. It is interesting to compare these results with earlier estimates of the effect of federal matching grants on welfare expenditures. Both Osman and Sacks and Harris estimated state per capita welfare expenditures as a function of per capita federal welfare grants and other independent variables. While both the dependent and independent variables are thus different from those used here, it can be shown that their coefficients are conceptually equivalent to the coefficient $a_5$ of the present model, and that the numerical values of the two coefficients are therefore directly comparable. Osman obtained an estimate of 1.37 for this coefficient, and Sacks and Harris found coefficients of 1.64 and 1.68, as compared to my estimates of 1.04 and 1.55, using the same (biased) approach. In both of the earlier studies, this result was interpreted as implying that federal grants have a "stimulative" effect on state and local expenditures. The empirical results obtained in this study indicate that these ordinary least squares results are almost entirely attributable to simultaneous equation bias. (See Edward Gramlich for an excellent summary of the extensive literature on federal grants-in-aid.)

[11] It should be noted that $\hat{F}$ is *not* simply a linear combination of the other independent variables in-

technique was used to obtain an estimated value of $MP$. The dependent variable $B$ was first regressed on all of the independent variables of equation (11) except $MP$ and $F$. Then, for all states using the original *AFDC* matching formula, the marginal matching rate associated with the estimated value of $B$ was derived from the matching formula. For these states, the actual value of $MP$ was replaced with this estimated value to form the variable $\widehat{MP}$. Use of $\hat{F}$ and $\widehat{MP}$ to estimate $a_4$ and $a_5$ eliminates any simultaneity bias in these coefficients.

Using this two-stage procedure, equation (11) was estimated separately for the years 1963–67 and 1968–72; the estimated coefficients (and associated statistics) are shown in the first two columns of Table 2. Aside from the regional dummies, only the coefficients of $\hat{F}$ in both periods, and the time variables in the later period, are insignificant at the 95 percent confidence level (one-tail test).

On the basis of these coefficients, we conclude that apart from the marginal price effect measured by $\widehat{MP}$, federal matching has no significant impact on state payment levels; i.e., federal funds simply displace state funds dollar for dollar. This finding, and the relatively modest size of the estimated coefficients of $\hat{F}$, stands in marked contrast with earlier published results on the effects of matching. (See fn. 10.) This should not be a surprising result. As the theoretical model indicates, matching only has an income effect on $B$ insofar as it lowers taxpayers' liabilities, raises their after-tax income, and thereby reduces their marginal utility of own-income. While the direction of the effect is clear, its magnitude is likely to be extremely small empirically. The average federal share of *AFDC* payments per recipient in this sample is \$272. Since the average recipient/population ratio is only about .025, this amounts to an increase in annual per capita after-tax income of only about \$7, as compared with an average per capita income of \$3000. One would not expect such a minor change in after-tax income to have much effect upon the marginal utility of own-income.

Because $\hat{F}$ and the time variables are correlated with several of the other independent variables, their inclusion in the regression equation may result in inefficient estimates of the other coefficients. Therefore, equation (2) was reestimated without these three variables and equation (1) was reestimated without $\hat{F}$; the resulting estimates are shown in columns (3) and (4) of Table 2. The discussion below is based on these estimates.

Although federal matching does not appear to have a significant income effect on benefit levels, the marginal price effect is significantly different from zero at the 99 percent confidence level. The coefficients of $\widehat{MP}$ imply an elasticity with respect to the price variable of about $-.23$ in both periods, evaluated at the point of means.[12] While the estimated elasticity is thus relatively small, these estimates imply that large changes in the matching formula can have nonnegligible effects upon payment levels. For example, suppose that the current *AFDC* matching system were replaced by a federally financed benefit floor with the states allowed to provide supplementary payments at their own expense. This

cluded in the regression, because in the first-stage regression the $X_i$ are multiplied by $m$, which varies widely across states. The $R^2$ of the first-stage regression was .68 in 1963–67 and .62 in 1968–72.

[12] It should be noted that these coefficients are subject to a possible bias. As noted above, since 1965 states have been allowed the option of reimbursement for *AFDC* expenditures at the matching rate established for the Medicaid program. Under the Medicaid matching formula, the marginal price is positively correlated with state per capita income. This may impart an upward bias if the income variable used here does not adequately control for income. Thus, the price effect of matching is, if anything, stronger than indicated by these coefficients.

TABLE 2—ESTIMATED REGRESSION COEFFICIENTS
(and $t$-statistics)

| Time Period | 1963–67 (1) | 1968–72 (2) | 1963–67 (3) | 1968–72 (4) | 1963–67 (5) | 1968–72 (6) |
|---|---|---|---|---|---|---|
| Constant | 1054. | 677. | 1512. | 663. | 1529. | 579. |
| $Y$ | .42 | .63 | .43 | .64 | .43 | .66 |
| | (6.95) | (9.59) | (7.15) | (11.47) | (6.42) | (9.98) |
| $RPOP$ | −8531. | −6521. | −9546. | −6905. | −9438. | −6661. |
| | (−3.46) | (−2.59) | (−4.12) | (−3.81) | (−3.74) | (−3.47) |
| $RCPT$ | .38 | .24 | .38 | .25 | .40 | .21 |
| | (2.06) | (2.06) | (2.05) | (2.34) | (2.11) | (1.86) |
| $\widehat{MP}$ | −466. | −655. | −455. | −672. | −462. | −691. |
| | (−4.25) | (−7.14) | (−4.34) | (−9.20) | (−4.36) | (−9.28) |
| $\hat{F}$ | .40 | .09 | | | | |
| | (1.11) | (0.33) | | | | |
| $NWH$ | −308. | −391. | −367. | −419. | −396. | −362. |
| | (−2.91) | (−2.53) | (−4.00) | (−3.52) | (−3.81) | (−2.71) |
| $NE$ | 96. | 143. | 81. | 148. | 77. | 149. |
| | (1.81) | (2.07) | (1.57) | (2.32) | (1.48) | (2.33) |
| $W$ | 18. | −99. | 22. | −102. | 20. | −98. |
| | (0.40) | (−1.66) | (0.48) | (−1.81) | (0.43) | (−1.74) |
| $OS$ | −568. | −664. | −685. | −690. | −764. | −536. |
| | (−4.28) | (−5.71) | (−8.47) | (−8.22) | (−3.85) | (−1.94) |
| $BS$ | −144. | −249. | −137. | −248. | −185. | −79. |
| | (−2.18) | (−3.06) | (−2.08) | (−3.19) | (−1.34) | (−0.48) |
| T | −148. | −27. | −143. | | −143. | |
| | (−2.04) | (−0.13) | (−1.97) | | (−1.96) | |
| T$SQ$ | 16. | 1. | 15. | | 15. | |
| | (1.75) | (0.12) | (1.66) | | (1.65) | |
| $NWHOS$ | | | | | 126. | −231. |
| | | | | | (0.40) | (−1.14) |
| $NWHBS$ | | | | | 91.0 | −315. |
| | | | | | (0.45) | (−1.23) |
| $R^2$ | .73 | .78 | .73 | .78 | .73 | .78 |
| $N$ | 255. | 255. | 255. | 255. | 255. | 255. |

*Sources of Data:*

$Y$: *Survey of Current Business*, Table 2, p. 43, and *Economic Report of the President*, Table C44.
$RCPT$, $B$: *Public Assistance Statistics*, Tables 4 and 7.
$RPOP$: *Current Population Reports* . . . , and *Public Assistance Statistics*, Tables 4 and 7.
$F$: *Public Assistance Statistics*, Tables 4 and 7; *Sources of Funds Expended for Public Assistance Payments*; *Funds, by Source, Expended for Public Assistance Payments*; and unpublished data for calendar year 1971 from the U.S. Department of Health, Education, and Welfare.
$MP$: *Characteristics of State Public Assistance Plans* and unpublished data from the U.S. Department of Health, Education, and Welfare.
$NWH$: *Findings of the 1971 AFDC Study*, Table 2; *Findings of the 1969 AFDC Study*, Table 2; *Findings of the 1967 AFDC Study*, Table 2; and *Characteristics of Families Receiving AFDC, 1961*, Table 3.
$OS$: Dummy = 1 for Alabama, Arkansas, Florida, Georgia, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Texas, Virginia.
$BS$: Dummy = 1 for Delaware, District of Columbia, Kentucky, Maryland, Oklahoma, West Virginia.
$NE$, $W$: Census region definitions.

amounts to elimination of federal matching at the margin. The results imply that in the 1968–72 period such a change would have reduced *AFDC* benefit levels by about 16 percent, or about $320 annually for a family of four, a reduction of over $.6 billion in total *AFDC* payments.

The results also indicate, not surpris-

ingly, that the level of transfers is strongly influenced by state per capita income, as predicted by the model. The coefficients of $Y$ indicate that a \$1.00 increase in real state per capita income raises real *AFDC* payments to a family of four by approximately \$.43 in the earlier period and \$.64 in the later period. The estimated elasticities of $B$ with respect to per capita income (at the point of means) on the basis of these coefficients are .65 and 1.17. The income elasticity of *AFDC* payments thus appears to have shifted upward somewhat over the period analyzed.

The predictions of the theoretical model are also supported by the highly significant coefficients of the recipient/population ratio and the absolute number of recipients, both of which have the expected sign in both periods. The significance of these variables is probably the most stringent test of the public goods nature of income redistribution under the *AFDC* program; failure to find significant effects for these variables would cast serious doubt on the applicability of the public goods model.[13]

Several observations are in order with respect to the coefficients of *RCPT* and *RPOP*. First, consider the effect of an increase in the *AFDC* caseload on benefit levels, holding state population constant. The direct effect of an increase in the number of recipients (*RCPT*) is positive, whereas the effect of the concomitant increase in the recipient/population ratio (*RPOP*) is negative. The net effect of these two opposing influences will depend on the level of state population. For example, differentiating the regression equation for 1968–72 with respect to *RCPT* yields:

$$\partial B/\partial RCPT = .25 - 6905./POP \tag{14}$$

where *POP* is state population in thousands. This expression is negative for population values less than 27,620,000. For 1963–67, the corresponding expression is negative for population values less than 25,121,000. Thus caseload increases tend to depress benefit levels in all states. However, the net effect approaches zero as state population approaches these critical values, so that this effect is greater in small states than in large ones.

Second, the estimated coefficients of *RPOP* are roughly consistent with those estimated for $\widehat{MP}$, the marginal matching rate. In the theoretical model developed in the previous section, the recipient/taxpayer ratio and the federal matching rate both enter the determination of benefits in the same manner, as multiplicative terms of the *mc* schedule. An equiproportional change in either variable will have the same effect on equilibrium benefit levels. Thus, the theoretical model predicts that the elasticities of benefit levels with respect to the two variables will be equal. On the basis of these estimates, the elasticity with respect to *RPOP* (at the point of means) is approximately $-.12$ in both time periods. The elasticity with respect to $\widehat{MP}$, $-.23$, is at least the same order of magnitude.

The racial composition of the recipient population also exerts a strong influence on the level of payments. The coefficients of *NWH* imply that, *ceteris paribus*, a family of four in a state with an all-nonwhite recipient population would receive about \$400 per year less than one in a state where all recipients were white.[14] This is evidence of rather substantial discrimina-

[13] The only other study, to my knowledge, to use either of these variables to explain welfare payment levels is by Gold, who used the recipient/population ratio in a regression analysis of total state public assistance payments per recipient in 1957 and 1966. He estimated elasticities of $-.07$ and $-.19$, as compared with the estimate of about $-.12$ found here for both five-year periods. Gold's estimated coefficient for 1957 was insignificant, while his coefficient for 1966 was significant at the 95 percent level (one-tail test).

[14] This is probably a conservative estimate of the effect of discrimination because as noted in fn. 9 above, *NWH* is subject to measurement errors which will bias its coefficient toward zero.

tion against nonwhite recipients on the part of (predominately white) taxpayers.[15] It should be noted that the racial composition variable is not simply acting as a proxy for low incomes or a political bias against welfare in southern states with predominately nonwhite caseloads, since both income and regional variables are explicitly included in the regression equation.[16] It is of some interest to inquire, however, whether the degree of discrimination against nonwhites is more severe in the South than in the rest of the nation—i.e., whether variations in the racial composition of the caseload have a greater effect on benefit levels in these states. To test this hypothesis, two interaction variables (*NWHOS* and *NWHBS*), formed by multiplying the Old South and border state dummies by *NWH*, were added to the regression equation. The results are shown in columns (5) and (6) of Table 2. In neither time period are these coefficients significantly different from zero. Thus, it appears that the degree of discrimination against nonwhites implicit in *AFDC* benefit levels does not differ significantly between the South and the rest of the nation.

While no signs were posited for the regional dummies, the estimated regional differentials conform roughly with what one might expect on the basis of casual observation. As compared with the omitted north central region, benefits are substantially (\$685–\$690) lower in the Old South and moderately (\$137–\$248) lower in the border states, *ceteris paribus*. There is also a small (\$102) negative differential in the western region in the later period and a modest (\$81–\$148) positive differential in the northeastern region. Although the dummy variables employed here may capture a variety of regional effects, it seems plausible that these differentials reflect systematic regional differences in taxpayers' preferences for income redistribution.

To provide an indication of the relative contributions of the independent variables to regional differentials in benefit levels, Table 3 presents mean benefits by region, regional differentials from the north central region mean, and the portion of regional differentials attributable to each variable. The effect of each variable on the regional differential in benefits is the product of the deviation of its mean in that region from the north central mean and the corresponding regression coefficient. Regional differentials are dominated by the intercept shift coefficients (i.e., the regional dummies). However, differences in mean marginal matching rates and per capita income also substantially affect regional variations in benefits, particularly in the South.[17]

## III. Conclusions

The empirical estimates presented here for the *AFDC* program over the ten-year period 1963–72 are remarkably consistent with the public good model of income re-

[15] Collins also found that the racial composition of the recipient population had a significant negative effect on *AFDC* payment levels in 1960. However, she does not report the regression coefficient for the race variable. The only other significant variable in her regression is state per capita income. Gold found highly significant effects of the racial composition of the state population on total public assistance payments per recipient in both 1957 and 1966. His estimated elasticities were −.14 and −.18, as compared with elasticities of −.09 and −.11 estimated here. Again, the only other variable which was significant in both years was state per capita income.

[16] Gold also found that the inclusion of a dummy variable identical to the *OS* dummy used here had no effect on the significant negative coefficients of his racial composition variables in 1957 and 1966. Surprisingly, however, the coefficient of the regional dummy itself was insignificant in both years.

[17] It should be noted that the differentials shown in Table 3 are not in any sense an absolute measure of the "importance" of each variable in determining benefit levels, since they are conditional upon the existing regional variation in the independent variables. In some cases (e.g., the marginal matching rate), this variation is susceptible to change by governmental policy, and in all cases the interstate variation is greater than the variation in regional means.

TABLE 3—REGIONAL DIFFERENTIALS IN *AFDC* BENEFITS

| | Old South | | Border States | | West | | Northeast | |
|---|---|---|---|---|---|---|---|---|
| | 1963–67 | 1968–72 | 1963–67 | 1968–72 | 1963–67 | 1968–72 | 1963–67 | 1968–72 |
| Mean Benefit (1967 dollars): | 933. | 1006. | 1480. | 1523. | 1742. | 1871. | 1885. | 2402. |
| Differential from north central mean: | −795. | −1024. | −247. | −507. | 14. | −159. | 166. | 372. |
| Differential due to differences in: | | | | | | | | |
| Intercept | −685. | −690. | −137. | −248. | 22. | −102. | 81. | 148. |
| *NWH* | −86. | −129. | −49. | −72. | 25. | 34. | 42. | 46. |
| *Y* | −265. | −349. | 26. | 71. | 9. | 12. | 79. | 143. |
| $\widehat{MP}$ | 285. | 196. | 75. | −116. | 4. | −54. | −20. | 78. |
| *RPOP* | −43. | −54. | −156. | −131. | −38. | −49. | −32. | −70. |
| *RCPT* | −1. | 2. | −6. | −10. | −8. | −1. | 16. | 27. |

distribution developed in Section I. State per capita income, the federal matching rate, the recipient/population ratio, the absolute number of recipients, and the racial composition of the caseload were all found to significantly affect state benefit levels in the directions predicted by the model. The significance of the recipient/population ratio and the absolute number of recipients is especially noteworthy, since these variables are particularly characteristic of the public good model.

The estimates also yield several important insights into the benefit structure of the state-run programs. The estimated elasticity of real *AFDC* benefits with respect to real state per capita income is close to unity, indicating that payment levels have kept pace with the growth of income. A more disturbing finding is the apparent discrimination against nonwhites in setting state benefit levels, which appears to exist in all regions. The estimates also reveal systematic regional differentials in *AFDC* benefits, particularly in the South, which may reflect fundamental differences in taxpayers' preferences for income redistribution.

The estimated coefficients of the recipient/population ratio and the absolute number of recipients indicate that the net effect of increases in the *AFDC* caseload is to depress benefit levels, *ceteris paribus*. The strength of this effect depends on population size, however, with larger effects occurring in the smaller states.

The empirical estimates of the effects of federal matching presented here have important policy implications. Previous studies of the effects of matching have been misspecified and have employed biased estimation techniques. When the separate price and income effects of matching are estimated, as required by the theory, it is seen that the income effect represented by the average federal share is not significantly different from zero. The price effect, represented by the marginal matching rate, is highly significant, however. Since a wide variety of possible federal welfare policies operate on these two variables independently, it is important to draw this distinction. For example, the institution of a federally financed benefit floor which eliminated federal matching at the margin while holding constant—or even increasing—the average federal share of benefits would reduce benefit levels by about 16 percent on the basis of these estimates.

## REFERENCES

**P. S. Albin** and **B. Stein**, "Determinants of Relief Policy at the Sub-Federal Level," *Southern Econ. J.*, Apr. 1971, *38*, 445–57.

G. S. Becker, *The Economics of Discrimination*, Chicago 1957.

D. Black, "On the Rationale of Group Decision Making," *J. Polit. Econ.*, Feb. 1948, *56*, 23–34.

D. F. Bradford and W. E. Oates, "Toward a Predictive Theory of Intergovernmental Grants," *Amer. Econ. Rev. Proc.*, May 1971, *61*, 440–48.

L. S. Collins, "Public Assistance Expenditures in the United States," in O. Eckstein, ed., *Studies in the Economics of Income Maintenance*, Washington 1967, 97–174.

S. Fabricant, *The Trend of Government Activity in the U.S. Since 1900*, New York 1952.

G. W. Fisher, "Determinants of State and Local Government Expenditures: A Preliminary Analysis," *Nat. Tax J.*, Dec 1961, *14*, 349–55.

———, "Public Assistance Expenditure," in *Report of The Commission on Revenue of the State of Illinois*, Springfield 1963.

———, "Interstate Variation in State and Local Government Expenditures," *Nat. Tax J.*, Mar. 1964, *17*, 57–74.

R. B. Gold, "Fiscal Capacities and Welfare Expenditures of States," *Nat. Tax J.*, Dec. 1969, *22*, 496–505.

R. B. Goldfarb, "Pareto Optimal Redistribution: Comment," *Amer. Econ. Rev.*, Dec. 1970, *60*, 994–96.

E. M. Gramlich, "The Effect of Federal Grants on State-Local Expenditures: A Review of the Econometric Literature," *Proc. Nat. Tax Assn.*, 1969, 569–93.

H. M. Hochman and J. D. Rodgers, "Pareto Optimal Redistribution," *Amer. Econ. Rev.*, Sept. 1969, *59*, 542–57.

——— and ———, "Redistribution and the Pareto Criterion," *Amer. Econ. Rev.*, Sept. 1974, *64*, 752–57.

E. O. Olsen, "A Normative Theory of Transfers," *Publ. Choice*, Spring 1969, *6*, 39–58.

J. W. Osman, "The Dual Impact of Federal Aid on State and Local Government Expenditures," *Nat. Tax J.*, Dec. 1966, *19*, 362–72.

M. V. Pauly, "Mixed Public and Private Financing of Education: Efficiency and Feasibility," *Amer. Econ. Rev.*, Mar. 1967, *57*, 120–30.

S. Sacks and R. Harris, "The Determinants of State and Local Government Expenditures and Governmental Flows of Funds," *Nat. Tax J.*, Mar. 1964, *17*, 74–85.

L. C. Thurow, "The Income Distribution as a Pure Public Good," *Quart. J. Econ.*, May 1971, *86*, 327–36.

G. Tullock, "The Charity of the Uncharitable," *Western Econ. J.*, Dec. 1971, *9*, 379–92.

G. M. von Furstenberg and D. C. Mueller, "The Pareto Optimal Approach to Income Redistribution: A Fiscal Application," *Amer. Econ. Rev.*, Sept. 1971, *61*, 628–37.

U.S. Bureau of the Census, *Current Population Reports, Population Estimates and Projections*, Series P-25, No. 460, June 1971, and No. 488, Washington, Sept. 1972.

U.S. Council of Economic Advisors, *Economic Report of the President*, Washington, Jan. 1973.

U.S. Department of Health, Education, and Welfare, *Characteristics of Families Receiving Aid to Families with Dependent Children, November-December, 1961*, Washington, Apr. 1963.

———, *Characteristics of State Public Assistance Plans Under the Social Security Act, General Provisions*, Washington annual reports, 1962–72.

———, *Findings of the 1967 AFDC Study, Part I*, Washington 1968.

———, *Findings of the 1969 AFDC Study, Part I*, Washington 1970.

———, *Findings of the 1971 AFDC Study, Part I*, Washington 1972.

———, *Funds, by Source, Expended for Public Assistance Payments and for Administration, Services, and Training, Calendar Year Ended December 31, 1972*, Washington 1973.

———, *Public Assistance Statistics*, Washington, Report A-2, June reports 1962-72.

———, *Source of Funds Expended for Public Assistance Payments*, Washington, Report F-1, annual reports for calendar years 1962–70.

*Survey of Current Business*, Washington, Aug. 1973.

# The Production Function and Technical Change in Postwar Soviet Industry: A Reexamination

By PADMA DESAI*

In a pioneering article in a recent issue of this *Review*, Martin Weitzman rigorously analyzed the phenomenon of postwar economic growth in Soviet industry. His paper represents the first systematic attempt at identifying the production function for Soviet industry without making the competitive relations assumption. It interprets the growth of Soviet industry in terms of the elasticity of substitution between capital and labor, and the rate of technical change.

Weitzman's main conclusions are the following: (i) the constant elasticity of substitution (*CES*) production function with constant returns to scale adequately represents postwar growth in Soviet industry; (ii) the estimated elasticity of substitution of .403 is significantly less than 1; and (iii) the estimated technical change parameter of .0205 is not different from the parameter value of .023 for American manufacturing in the recent period.

Given these parametric values, Weitzman reaches two major conclusions with respect to the actual performance of Soviet industry in the postwar period and its future growth. With respect to its actual performance, he concludes that by the late 1960's " . . . a low elasticity of substitution seems to imply that capital accumulation has outstripped labor growth by a wide enough margin that the drag due to diminishing returns is significantly cutting into output growth. In the mature Soviet economy, labor scarcity appears to be a reality" (p. 685). Furthermore, the rate of growth of technical change of about 2 percent per year " . . . is not nearly so significant a determinant of economic growth as in some other economies. The Soviet postwar record looks very much like a classical model of economic growth with high rates of capital accumulation serving as the prime mover" (p. 685).

As for the future implications of this scenario for Soviet industry, Weitzman concludes that "Instead of capital, labor and technical change will have to be increasingly relied upon as alternative sources of future economic growth" (p. 685). But since "the growth of industrial labor force cannot in the near future rise significantly above what it has averaged in the past five years" (pp. 685–86), the Soviets can achieve high rates of industrial growth by actively encouraging the residual element or technical change. "With in-

* Russian Research Center, Harvard University. In the course of working on this paper, I have profited from prolonged and frequent discussions with T. N. Srinivasan and valuable suggestions from Martin Weitzman. Useful comments were made on a penultimate draft by Zvi Griliches, Bent Hansen, Alan Manne, Thomas Marschak, Tom Rothenberg, and Robert Solow. Thanks are also due to Abram Bergson and Gregory Grossman for comments on an earlier version of the paper. The computer work involved in the paper was done with speed and expertise by Bronwyn Hall. The non-linear estimation program used has been developed by Robert Hall. The financial support provided by the Institute of International Studies at the University of California, Berkeley, during 1973–74, is gratefully acknowledged. An earlier version of this paper was discussed at the Stockholm School of Economics and presented at the Econometric Society meetings in San Francisco in December 1974.

creased λ [technical change], the Soviet growth scenario would resemble a little more closely the contemporary expansion of Japan and a few Western countries" (p. 686).

These results and conclusions are impressive enough to raise the critical question: how dependable are Weitzman's data, his specification of the production relationship for Soviet industry, and its theoretical underpinnings?

Weitzman relies exclusively on estimates of output, industrial productive fixed assets, and industrial man-hours available in Soviet sources. However, in the absence of corresponding data on raw materials, he is constrained to develop a "hybrid" output series, weighting the output indexes of Soviet branches by their estimated 1960 value-added shares, and to work with this hybrid series and the available capital and labor data.

Weitzman's hybrid series cannot really be interpreted as a value-added series. Weitzman himself is aware of the limitations of his hybrid index. Thus he writes:

> In pretending that industrial output is a function of capital and labor alone we are introducing an error by omitting industrial purchases of inputs from other sectors, principally transportation and agriculture. My distinct feeling is that such omissions turn out to be empirically irrelevant for industry as a whole. In any case it is difficult to obtain year to year input-output type data of the kind needed to test the importance of this omission. [p. 692]

The principal task of this paper, then, is to test the importance of this omission. In doing this, I have developed a data series of raw materials used in Soviet industry, utilizing a technique which has been set out at length elsewhere,[1] and which I sketch briefly in the Statistical Appendix. Section A of the Appendix shows how gross output, value-added, and raw materials in Soviet industry at current prices can be derived. The former are described as Desai estimates as they differ from the official Soviet indexes of gross output.[2] Since I need these indexes at constant prices, the necessary price deflators are developed in Section B of the Appendix.

The resulting Desai estimates of gross output, raw materials, and gross value indexes are presented alongside and compared with the Soviet output index and Weitzman's hybrid index in Section I. Next, in Section II, alternative production functions are specified for Soviet industry, using both gross value-added (Desai) and gross output (Soviet and Desai) as the dependent variables. While the Weitzman assumptions of Hicks-neutral, disembodied technical change, and a multiplicative error term are retained, and his use of nonlinear estimation technique is retained in preference to the use of competitive relations for estimating the production function, my specification of the production function in terms of gross output will allow for raw materials.

Section III presents the "best" estimates. Among them, in turn, the *CES* production function with constant returns to scale, *and* with raw materials specified, appears to give the best fit. While the elasticity of substitution between capital and labor continues to be significantly less than one, the Hicks-neutral technical change parameter increases substantially to 3.5 to 4 percent annually. This and other implications of the analysis, plus directions for further econometric analysis, conclude the paper.

## I. Statistical Series

The statistical methods set out in the

[1] See my 1973 paper.

[2] For various cross-checks on my estimates, see my 1973 paper.

TABLE 1—INDEXES OF GROSS OUTPUT, GROSS VALUE-ADDED, AND MATERIAL PURCHASES IN SOVIET INDUSTRY

| | Gross Value-Added | | | | Material Purchases | | Gross Output (Desai estimate) | | Soviet Gross Output Index | | Weitzman "Hybrid" Index | | Desai "Hybrid" Index | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Year | Single Deflation (1) | Rate of Growth of (1) (2) | Double Deflation (3) | Rate of Growth of (3) (4) | Index (5) | Rate of Growth of (5) (6) | Index (7) | Rate of Growth of (7) (8) | Index (9) | Rate of Growth of (9) (10) | Index (11) | Rate of Growth of (11) (12) | Index (13) | Rate of Growth of (13) (14) |
| 1955 | 57.93 | | 59.07 | | 56.30 | | 57.22 | | 61.07 | | 60.34 | | 61.20 | |
| 1956 | 62.49 | 7.9 | 67.40 | 14.1 | 63.05 | 12.0 | 64.50 | 12.7 | 67.43 | 10.4 | 66.33 | 9.9 | 67.25 | 9.9 |
| 1957 | 71.78 | 14.9 | 76.06 | 12.8 | 69.20 | 9.7 | 71.49 | 10.8 | 74.34 | 10.2 | 73.01 | 10.1 | 73.86 | 9.8 |
| 1958 | 80.93 | 12.7 | 83.63 | 9.9 | 77.69 | 12.3 | 79.67 | 11.4 | 81.91 | 10.2 | 81.29 | 11.3 | 81.83 | 10.8 |
| 1959 | 86.78 | 7.2 | 91.38 | 9.3 | 82.65 | 6.4 | 85.56 | 7.4 | 91.12 | 11.2 | 90.56 | 11.4 | 89.01 | 8.8 |
| 1960 | 100.00 | 15.2 | 100.00 | 9.4 | 100.00 | 21.0 | 100.00 | 16.9 | 100.00 | 9.7 | 100.00 | 10.4 | 100.00 | 12.3 |
| 1961 | 115.91 | 15.9 | 116.11 | 16.1 | 120.39 | 20.4 | 118.96 | 19.0 | 109.00 | 9.0 | 109.56 | 9.6 | 109.31 | 9.3 |
| 1962 | 125.97 | 8.7 | 129.12 | 11.2 | 126.54 | 5.1 | 127.40 | 7.1 | 120.00 | 10.1 | 120.27 | 9.8 | 119.90 | 9.7 |
| 1963 | 142.19 | 12.9 | 152.27 | 17.9 | 138.38 | 9.4 | 143.01 | 12.2 | 129.00 | 7.5 | 131.21 | 9.1 | 130.37 | 8.7 |
| 1964 | 151.46 | 6.5 | 147.04 | −3.4 | 148.93 | 7.6 | 148.30 | 3.7 | 139.00 | 7.5 | 141.78 | 8.1 | 140.26 | 7.6 |
| 1965 | 164.49 | 8.6 | 177.30 | 20.6 | 157.72 | 5.9 | 164.25 | 10.8 | 151.00 | 8.6 | 153.41 | 8.2 | 151.88 | 8.3 |
| 1966 | 175.42 | 6.6 | 178.93 | .09 | 177.15 | 12.3 | 177.74 | 8.2 | 164.00 | 8.6 | 167.09 | 8.9 | 165.28 | 8.8 |
| 1967 | 188.07 | 7.2 | 179.03 | — | 185.69 | 4.8 | 183.39 | 3.2 | 180.00 | 9.8 | 183.65 | 9.9 | 181.78 | 10.0 |
| 1968 | 211.43 | 12.4 | 204.50 | 14.2 | 206.58 | 11.2 | 205.88 | 12.3 | 195.00 | 8.3 | 199.91 | 8.9 | 197.91 | 8.9 |
| 1969 | 227.56 | 7.6 | 227.97 | 11.5 | 206.94 | — | 213.95 | 3.9 | 209.00 | 7.2 | 215.2 | 7.6 | | |
| 1970 | 251.21 | 10.4 | 267.36 | 17.3 | 217.00 | 4.9 | 233.79 | 9.3 | 227.00 | 8.6 | | | | |
| 1971 | 263.23 | 4.8 | 289.34 | 8.2 | 222.12 | 2.4 | 244.52 | 4.6 | 244.66 | 7.8 | | | | |

*Notes:* All figures in columns (1)–(8) and (13)–(14) are derived according to the method outlined in the Appendix. Figures in columns (9)–(10) are directly from Soviet sources. Figures in columns (11)–(12) are from Weitzman (p. 677). All figures in columns (2), (4), (6), (8), (10), (12), and (14) are the annual percentage growth rates of the corresponding figures in the preceding column.

Appendix result in the following (constant-price) indexes of importance in our estimation of a production function for Soviet industry: gross value-added; material usage; and gross output (Desai). These are listed in Table 1, together with the following additional indexes: Soviet gross output;[3] Weitzman's "hybrid";[4] and Desai's "hybrid" (using Weitzman's weighting procedure but Desai's estimated value-added weights).[5] The annual percentage rates of growth of each series are also presented in Table 1. A few remarks on the different series are in order before the estimation of production functions, based on these series, is described in the following sections.

It is clear now that the absolute levels of the indexes estimated by my methods (and presented in columns (1), (3), and (7)) are generally, but especially after 1960, higher than the indexes in columns (9), (11), and (13). The gross value-added index I estimated *via* the double deflation method is generally higher than all the alternative index series.[6] Secondly, the gross output index I estimated (column (7) of Table 1) and the index of gross output in Soviet sources (column (9) of Table 1) are not far apart, although the corresponding rates of growth show consider-

[3] This series is taken from the following sources: For 1955–57, *Promyshlennost' S.S.S.R.*, p. 36; for 1958–60, *Narkhoz* (1965, p. 125); for 1960–64, *Narkhoz* (1965, p. 126); for 1965–69, *Narkhoz* (1967 and 1969, pp. 189 and 147); for 1970, *Narkhoz* (1970, p. 136); for 1971, *Jubilee Volume*, p. 126.

[4] This series is taken from Weitzman, p. 677.

[5] The value-added figures in billion rubles, the corresponding Desai weights, and the Weitzman weights, p. 690, are shown in the Appendix, Table A3.

[6] Also the double-deflated value-added series show a rising trend with marked fluctuations. In my judgment, the fluctuations are a result of the method employed in deriving the series. On the other hand, they may reflect changes in inventories of materials. There is no way of tracking down these changes, if any, and adjusting for them on the basis of evidence on inventories in Soviet sources. Furthermore, a dip in the double-deflated value-added index in 1964 also seems to arise from the use of "inappropriate" price deflators for gross output and/or raw materials. Paul David (1962) has suggested that such dips can arise as a result of a change in the technique of production. Such a possibility, which is likely in the case of double-deflated value-added index number comparisons between two points of time which are far wide apart, can be ruled out in the present case where the comparison is between two consecutive years.

able divergence from year to year.[7]

The most striking observation is with respect to the Soviet gross output index (column (9)), the Weitzman hybrid index (column (11)), and the Desai hybrid index (column (13)). These three indexes and their growth rates are practically identical. The Weitzman index and the Desai index are identical, thus bearing out Weitzman's argument that "the industrial production index appears to be remarkably insensitive to changes in the value-added weights caused by somewhat different assumptions used in composing them" (p. 691).[8] Finally, in the estimation that follows in the next two sections, I used the Weitzman estimates of industrial manhours and his series on fixed assets to focus sharply on the differences introduced by my explicit consideration of raw materials in contrast to Weitzman's analysis.

## II. Specification of the Production Function

The specification of the production function and the underlying assumptions in the following analysis are essentially similar to those of Weitzman. The difference is that the dependent variable is defined here alternately in terms of gross output and gross value-added, and raw materials are explicitly considered. By contrast, Weitzman's dependent variable was a hybrid index (an overall industry output index where the branch *output* indexes are weighted by value-added-share weights), and raw materials were ignored among the explanatory variables. Furthermore, as noted in the previous section, Weitzman's hybrid index is almost identical to the Soviet gross output index for total industry. This being so, the omission of raw materials as one of the explanatory variables in the production relationship should distort Weitzman's parametric estimates.

The *CES* and Cobb-Douglas forms with variable and constant returns to scale were fitted to the time-series data. Two gross output series, Soviet and Desai, and Weitzman's hybrid series,[9] were used for estimating production functions defined on capital, labor, *and* raw materials. The two value-added series I derived, double-deflated and single-deflated, were used to estimate production functions defined on capital and labor alone.[10] Thus, given that

[7] The main reason for the divergence is that whereas I derive the gross output series first in current prices and then in constant prices by deflating the former with a suitable price index, the Soviet method is to derive the output series, beginning with the enterprise, by multiplying physical enterprise output with the product price of a given year. Incidentally, the average rates of growth of value-added (single deflated and double deflated) in columns (2) and (4) are higher than the average rates of growth of gross output, Desai and Soviet, in columns (8) and (10). The higher rate of growth of value-added in Soviet industry (with 1950 = 100 and for a slightly shorter period) as compared to that of gross output is also emphasized by Judith Thornton (1970).

[8] In my judgment, the industrial production index would remain "insensitive" not only because the value-added weights are spread over a range of .66 percent (for Desai weights) and .82 percent (for Weitzman weights) in Glass and ceramics and of 29 percent for both in Machinebuilding and metalworking but *also* because the spread in the branch *production* indexes themselves (to which the weights are applied), at least up to 1963, is not very large. For example, in 1963, the minimum production index is 109.64 in Soft goods and the maximum is 149.72 in Machinebuilding and metalworking, thus indicating a spread of 40.08 points.

[9] Weitzman's hybrid index may be treated as if it was an output index because it works with Soviet branch output indexes, merely using value-added for weights. Strictly speaking, the Weitzman index can be interpreted as a value-added index if materials are a fixed proportion of gross output in each branch throughout the period under consideration. This is, however, an extremely restrictive assumption and is not validated by an examination of the gross output and materials series employed in our analysis.

[10] It is relevant here to indicate the theoretical rationale for adopting either the single-deflated or the double-deflated value-added series in specifying the production function. David (1966) has shown that if observed payments to factors (i.e., value-added) result from profit maximization subject to the constraint of a linear homogeneous production function, under conditions of perfect competition in product and material markets, then the net output index can be obtained simply by deflating an index of current value-added by the index of the industry's product price. David thus provides a rationalization for using a single-deflated value-added index in the specification of a production function. He further argues that the double deflation

(in a three-factor production function) each *CES* form was tried with four alternative parametric restrictions on pairwise elasticities of substitution,[11] altogether 30 trials were made for the gross output production function and 8 for the value-added variety.

Of these 38 trials, the presentable results were (fortunately) confined to 5. All 5 were *CES*, none Cobb-Douglas, and none showed variable returns to scale. In essence, then, the following three forms of the *CES* production function were fitted with (varying degrees of) success:

(1) *CES (I): Constant returns to scale*

$$O_I(t) = A_I e^{\lambda_I t}[\alpha_I K_I(t)^{-\rho_I} + \beta_I L_I(t)^{-\rho_I} + (1 - \alpha_I - \beta_I) R_I(t)^{-\rho_I}]^{-1/\rho_I} e^{u_I(t)}$$

where

$$A_I > 0, \quad 0 \leq \alpha_I + \beta_I \leq 1, \quad \alpha_I \geq 0, \quad \beta_I \geq 0, \quad \rho_I \geq -1$$

and elasticity of substitution between

$$[K_I(t), L_I(t)] = [K_I(t), R_I(t)] = [L_I(t), R_I(t)] = 1/\rho_I + 1$$

(2) *CES (II): Constant returns to scale*

$$O_I(t) = A_I e^{\lambda_I t}[\alpha_I K_I(t)^{-\rho_I} + (1 - \alpha_I) L_I(t)^{-\rho_I}]^{-\gamma_I/\rho_I} R_I(t)^{(1-\gamma_I)} e^{u_I(t)}$$

where

$$A_I > 0, \ 0 \leq \alpha_I \leq 1, \ 0 \leq \gamma_I \leq 1, \ \rho_I \geq -1$$

and elasticity of substitution between

$$[K_I(t), L_I(t)] = \frac{1}{\rho_I + 1}$$

$$[K_I(t), R_I(t)] = 1$$

$$[L_I(t), R_I(t)] = 1$$

(3) *CES (I): Constant returns to scale*

$$Y_I(t) = A_I e^{\lambda_I t}[\alpha_I K_I(t)^{-\rho_I} + (1 - \alpha_I) L_I(t)^{-\rho_I}]^{-1/\rho_I} e^{u_I(t)}$$

where

$$A_I > 0, \quad 0 \leq \alpha_I \leq 1, \quad \rho_I \geq -1$$

In specifications (1) and (2), $O_I$ refers to gross output, and in specification (3), $Y_I$ refers to gross value-added. In all the specifications, $K_I$, $L_I$, and $R_I$ refer respectively to fixed productive assets, annual man-hours, and material usage.

## III. Results and Conclusions

### A. *"Best" Estimates*

In Table 2 the estimated parameters for trials, corresponding to the three production function forms just listed and corresponding to alternative estimates of the dependent variable, have been presented. All the parametric values satisfy the specified limits; but not all of them are statistically significant.

While some readers will find it possible to accept more than one of the five results presented in Table 2, my view is that the best result is for *CES* with gross output and identical elasticities of substitution

method requires no formal assumption, and "... because it makes no theoretical assumptions, it yields no more than hypothetical calculations which are not particularly appropriate for production function analysis" (p. 425). However, a rationale can be provided for the use of double-deflated value-added on the basis of further work by Kenneth Arrow and Bent Hansen. Thus Arrow has argued that if capital and labor are (weakly) separable from materials in production of gross output and if both the gross output and the "nested" production functions are homogeneous, then with further assumptions of profit maximization, it is possible to derive an index number of value-added up to a proportionality factor. Carrying the argument further, Hansen has argued that under the same assumptions, the Arrow index of value-added can be interpreted as a double-deflated value-added index. Hansen further argues that David's single-deflated value-added index differs from a double-deflated index by a term which can be interpreted as a terms-of-trade term (p. 10). It would, therefore, seem that under the assumptions of linear homogeneous production functions, profit maximization and separability, both the single-deflated and double-deflated value-added index series have a theoretical rationale for production function analysis.

[11] Thus, in the first variation of the *CES* form, the elasticity of substitution among the three factors taken two at a time is defined to be equal to, say, $\sigma$; in the remaining three variations, the elasticity of substitution between a given pair of factors is defined to be equal to $\sigma$, whereas the elasticity between the remaining two pairs of factors is specified to be equal to unity.

TABLE 2—PRODUCTION FUNCTION ESTIMATES FOR SOVIET INDUSTRY: 1955–1969[a]

| Form of equation (1)[b] | Dependent variable (2)[c] | Hicks-neutral technical change parameter (3)[d] | Elasticity of substitution[e] (4) | $\alpha$ (Capital coefficient) (5) | $\beta$ (Labor coefficient) (6) | $\gamma$ (7) | Returns to scale (8) | $R^2$ (9)[f] | $d$ (10)[f] | $SSR$ (11)[g] |
|---|---|---|---|---|---|---|---|---|---|---|
| (1) *CES (I)* | Soviet gross output | 0.0408 | 0.2771 | 0.0210 | 0.6798 | 0.2992 | | 0.9994 | 1.8591 | 0.0012 |
| | | (0.0052) | | (0.1811) | (0.0857) | | | | | (0.0112) |
| (2) *CES (II)* | Soviet gross output | 0.0354 | 0.1987 | 0.1997 | 0.8003 | 0.7611 | | 0.9994 | 1.6154 | 0.0013 |
| | | (0.0031) | | (0.1361) | | (0.0984) | | | | (0.0112) |
| (3) *CES (II)* | Desai gross output | 0.0034 | 0.1758 | 0.8801 | 0.1199 | 0.3184 | | 0.9990 | 2.3292 | 0.0026 |
| | | (0.0071) | | (0.2862) | | (0.1295) | | | | (0.0162) |
| (4) *CES (I)* | Single-deflated value-added | 0.0152 | 0.2070 | 0.8619 | 0.1381 | — | | 0.9970 | 1.5629 | 0.0084 |
| | | (0.0152) | | (0.1837) | | | | | | (0.0276) |
| (5) *CES (I)* | Double-deflated value-added | 0.0101 | 0.1686 | 0.8930 | 0.1070 | — | | 0.9913 | 2.3144 | 0.0222 |
| | | (0.0187) | | (0.2270) | | | | | | (0.0449) |

[a] All the values in parentheses are the standard errors of estimates of the relevant coefficients. The standard errors of estimates of the elasticity of substitution coefficient $\sigma$ (where $\sigma=1/1+\rho$) have not been stated. The $t$-values of the estimated coefficients $\rho$ in each case indicate that at the 95 percent confidence interval, the values of the elasticities of substitution coefficients are statistically significant.

[b] All production functions are characterized by constant returns to scale. The forms of the equation in column (1) are spelled out in detail in the text.

[c] The dependent variables in column (2) in terms of indexes with 1960=100 are stated in Table 1.

[d] The equations in the text also specify the limits to the values of the parameters in columns (3) to (8).

[e] The elasticity of substitution in *CES (I)* is equal at $1/1+\rho$ for all pairs of factors. For *CES (II)*, the elasticity reported in the column is between capital and labor, whereas that for all other pairs is equal to unity.

[f] $R^2$ in column (9) is (correlation coefficients)$^2$, $d$ in column (10) is the Durbin-Watson statistic.

[g] $SSR$ in column (11) is the sum of squared residuals and the values in parentheses are the standard errors of the regression.

between each pair of factors in row (1). This can be argued as follows: since the $R^2$ is high and almost equally good in all cases, we can shift attention to the sum of squared residuals in column (11). It is seen that the last three production functions show a substantially higher sum of squared residuals; hence the first two, *CES* on gross output, would seem more acceptable on that account. Note also that for the last three equations, the standard errors of estimate for the technical change parameters are too high and make them statistically insignificant.

Hence the (first) two *CES* production functions with alternative specifications of the factor elasticities of substitution seem to be the best fits. However, among the two, we may further prefer the first form, with the elasticity of substitution between each pair of factors at approximately 0.28, over the second form which postulates an elasticity of substitution between capital and materials, and between labor and materials, at unity and thus in excess of the estimated elasticity of substitution between capital and labor only at approximately 0.20.

### B. *Economic Implications*

Regardless, however, which of the two *CES* forms on gross output is chosen as the more acceptable, my conclusions are similar to Weitzman's in two respects: the *CES* constant returns to scale production function adequately describes the production relationship in Soviet industry; and the elasticity of substitution between capital and labor is significantly less than one, thus indicating diminishing returns to factor proportions. An amusing consequence of this finding by Weitzman and myself is that the *imputed* (competitive) share of labor income in Soviet industry *rises* over time whereas the *actual* (Soviet) share of labor income has *fallen* over time. The Marxist-Leninist system would thus seem to be of dubious trendwise value to the industrial proletariat compared to labor's potential share under the capitalist-

competitive system![12]

However, in *contrast* to Weitzman's estimated Hicks-neutral technical change parameter of 2 percent, my analysis leads to a higher estimate of 3.5 or 4 percent. This finding is clearly of some importance in suggesting that the rate of Soviet technical change is significantly higher than that arrived at by Weitzman. The conclusion that the Soviet technical change is as high as 3.5 or 4 percent is, of course, fully consistent with the widespread view of the generally low *level* of Soviet technology and, in fact, may even complement that view if one attributes the high *rate* of technical change to a process of "catching up."[13]

[12] The difference between imputed and actual shares is referred to by Thornton (Section IV, in her important paper). However, she argues: "But the Cobb-Douglas form with fixed factor shares provides a poor approximation to the data for Soviet industry in this period when the share of capital shows a considerable increase," and "In his production function study, Weitzman found elasticities of substitution ranging in value from .274 to .403. In contrast, the data presented in this study in which a rise in the capital-labor ratio is associated with an increasing share of capital would seem to imply an elasticity of substitution greater than unity" (pp. 867–68). The imputed and actual shares can in this manner be discussed interchangeably only if the Soviet system has worked within a fully competitive framework during the period under consideration. This, however, would be an extraordinary assumption to make. It is precisely for this reason that both Weitzman and I have used the non-linear estimation technique for estimating the production function for Soviet industry and avoided using the competitive-relations estimation technique pioneered by Arrow et al.

[13] I might also add that, in a separate study in which I analyze the forms of the production function, the elasticities of substitution, the returns to scale, and the technical change parameters for various *branches* of Soviet industry in the postwar period (using identical estimation procedures to those used in this paper), I get the following values of the technical change parameters: Electricity: 1.69–3.15 percent; Ferrous metallurgy: 1.74 percent; Chemical and mining-chemical: 7.68 percent; Machinery and metalworking: 4.89 percent; Construction materials: 2.80–3.36 percent; Light industry: 2.16 percent; Food industry (i.e., mainly processed food items): 4.96 percent. Leaving out Fuel, Wood, woodworking and paper-cellulose, and Glass and ceramics, these branches nonetheless cover a preponderant amount of Soviet industrial output. The approximate average of these branch technical change parameters indicates an aggregate industry parameter closer to 3.5 to 4 percent than Weitzman's 2 percent.

However, a further comment on the contrast between the estimates by Weitzman and myself is in order. The higher estimate in my analysis which accounts for the material inputs omitted in Weitzman's analysis could well indicate that technical change in Soviet industry was material saving.[14] Since the estimated elasticity of factor substitution is also less than one, it may well be that an examination of factor-augmenting forms for the technical progress function may yield results that indicate biased technical change.[15] Since the coefficient of capital stock in my preferred function is statistically insignificant, it is also possible that the introduction of biased technical progress could change the situation in that regard as well. This is clearly the next analytical step to be taken in the interpretation of technical progress in Soviet industry.

## Statistical Appendix

### A. *Method of Deriving Gross Value of Output, Gross Value-Added, and Material Usage in Soviet Industry at Current Prices*

The method of deriving these three components in Soviet industry is stated at length elsewhere.[16] Here I briefly recapitulate the essential steps. Note first that:

$$D_I(t) = d_I(t) \cdot K_I(t) \tag{A1}$$

[14] This would also seem to be indicated by the fact that the technical change parameters for *CES* on value-added (double-deflated and single-deflated) are substantially lower (even below Weitzman's) while being admittedly statistically insignificant. The hypothesis of material-saving technical progress in Soviet industry also sound *prima facie* plausible to those who think, à la Kennedy-Weizsacker lines, that the scarcity and/or dearness of a factor tends to bias technical change in the direction of seeking new techniques to economize on its use. Clearly Soviet firms, like firms in developing countries with quantitative restrictions on imports of materials, tend to carry high inventories and may also be directing their technical *R* and *D* efforts toward "reducing the dependence" on the precarious supply of materials in a centrally planned economy.

[15] Recall that both Weitzman and I use the Hicks-neutral form of technical progress in our estimates.

[16] See my 1973 article. It also extends the method to a derivation of these components in various branches of Soviet industry and further breaks down the estimates

where $K_I(t)$ is defined as the average productive fixed assets in industry as a whole during year t in terms of 1955 prices, $d_I(t)$ is the rate of amortization in industry, defined as a ratio of $K_I(t)$, and $D_I(t)$ is the level of depreciation or amortization in industry during year t.

The second set of relationships below links $D_I(t)$ thus obtained to total expenditure in industry during t:

$$E_I(t) = \frac{D_I(t)}{m_I(t)} \tag{A2}$$

where $E_I(t)$ is defined as *total* expenditure during t in industry on raw materials, fuel and electricity, amortization, wages plus social security deductions and "other expenditures," and $m_I(t)$ are the parameters which express amortization during t in industry as a proportion of *total* expenditure in industry.

The *total* expenditures $E_I(t)$ derived in (A2) above are in current enterprise prices, i.e., these prices are net of taxes; however, these expenditures include the taxes on purchases (such as oil and raw materials of Light and Food industries) made by industry. Next, we define gross output:

$$O_I(t) = E_I(t) + P_I(t) \tag{A3}$$

where $P_I(t)$ are the levels of industry profits given in Soviet sources. Moreover we have the definition of value-added:

$$Y_I(t) = P_I(t) + D_I(t) + W_I(t) \tag{A4}$$

where $W_I(t)$ are wages in total industry and are derived below:

$$W_I(t) = w_I(t) \cdot E_I(t) \tag{A5}$$

and $w_I(t)$ are available in Soviet sources.

It is also clear from equation (A2) above that the method outlined here also gives expenditures on materials (including auxiliary materials), fuel and electricity in industry. We hereafter refer to these as material purchases. Note that according to Soviet practice, these material purchases include purchases from sectors such as Transport and Communications.[17]

## B. *Method of Converting Current-Price into Constant-Price Estimates*

The next task consists of converting these three series of gross output, gross value-added, and material purchases at *current* factor-cost prices into constant prices with 1960 = 100. The gross output series are deflated with the Soviet price index in industry and then converted into indexes with 1960 = 100. The resulting estimates are given in column (7) of Table 1.

As for gross value-added, the standard procedure for converting gross *value-added* series in Soviet industry into constant prices has been to use the single deflation method: i.e., to employ price indexes of gross *output*.[18] Adopting this procedure, the resulting index of this series with 1960 = 100 is presented in column (1) of Table 1.

However, we can also use the double deflation method by (a) first converting both the gross output and material purchases series into constant prices with their respective price deflators, and (b) then subtracting the material purchases series from the gross output series. The method of deriving a suitable price index for deflating the material purchases in industry, then, consists of the following three steps:

(i) On the basis of the information in the 1959 and 1966 input-output tables given in Soviet sources,[19] total material purchases

of gross output in total industry between Groups A and B (i.e., Heavy and Consumer Goods sectors in Soviet terminology). The total industry, branch, and group estimates are compared for reliability with the corresponding figures in Soviet sources. All the sources of data derived from applying the set of equations in the Appendix, their specific features and their limitations are discussed at length in the article, and are not recapitulated here.

[17] Both the input-output tables for 1959 (*Narkhoz* 1960, pp. 104–43) and for 1966 (*Narkhoz* 1967, pp. 64–111) include purchases by various sectors of industry from Transport and Communications, Trade and Supplies and Others.

[18] For example, see Thornton, pp. 865–66.

[19] See *Narkhoz* (1960, pp. 104–43) and *Narkhoz* (1967, pp. 64–111). These tables indicate that purchases by industry from sectors such as Agriculture, Transport and Communications, Supplies and Others are respectively 36 percent and 37 percent of total material purchases by industry in 1959 and 1966. In my judgment, these figures of extra-industry purchases by industry are substantial and contradict Weitzman's assumption

TABLE A1—ABSOLUTE PRICES AND PRICE INDEXES IN SOVIET AGRICULTURE; TRANSPORT AND COMMUNICATIONS[a]

| (i) | Price Per Tonne of Agricultural Output (ii) | Price Index in Agriculture (iii) | Price Per Tonne/km. of Transport and Communications Output (iv) | Price Index in Transport and Communications (v) |
|---|---|---|---|---|
| 1958 | 144.299 | 107.10 | .598 | 86.79 |
| 1959 | 153.273 | 113.76 | .639 | 92.74 |
| 1960 | 134.736 | 100.00 | .689 | 100.00 |
| 1961 | 142.847 | 106.02 | .678 | 98.40 |
| 1962 | 164.962 | 122.43 | .662 | 96.08 |
| 1963 | 182.350 | 135.34 | .647 | 93.90 |
| 1964 | 148.581 | 110.27 | .646 | 93.76 |
| 1965 | 180.112 | 133.68 | .651 | 94.48 |
| 1966 | 177.536 | 131.76 | .651 | 94.48 |
| 1967 | 177.690 | 131.88 | .658 | 95.50 |
| 1968 | 176.697 | 131.14 | .643 | 93.32 |
| 1969 | 195.387 | 145.01 | .643 | 93.32 |
| 1970 | 211.650 | 157.08 | .653 | 94.71 |
| 1971 | 226.360 | 168.00 | .660 | 95.79 |

[a] Column (ii) is shown in rubles, and column (iv) in kopecks.

*Notes:* Gross values of agricultural and transport and communications outputs for 1958, 1963, 1963, and 1964 are taken from *Narkhoz* (1964, p. 677 and p. 67); for 1960 and 1965–70, from *Narkhoz* (1970, p. 607 and p. 60); for 1971, from *Jubilee Volume*, p. 59.

For 1959, 1961, and 1962, the gross values of outputs in each sector are derived by applying the proportions of these for each year to the absolute levels of gross outputs in the economy as a whole. The relevant information is given in *Narkhoz* (1962, p. 64).

Physical outputs in Agriculture for 1958–70 are taken from *Narkhoz* (1970, p. 273); for 1971, from *Jubilee Volume*, p. 217. Physical outputs in Transport and Communications for 1960–71 are from *Jubilee* Volume, p. 293; for 1958 and 1959, from *Narkhoz* (1965, p. 457).

Columns (ii) and (iv) are derived by dividing the gross values of outputs in each sector by the corresponding physical volumes of outputs.

Columns (iii) and (iv) are derived from the corresponding entries in columns (ii) and (iv) with 1960=100.

by industry are allocated in the proportions of 0.68, 0.21, and 0.11 to industry itself, to Agriculture, Transportation and Communications, respectively.

(ii) The price indexes in Soviet Agriculture, Transport and Communications are derived by evaluating the gross output in each at current prices and dividing it by the physical volume of the corresponding output. The series of absolute prices and the corresponding price indexes in Agriculture, Transport and Communications derived from the use of the relevant information in Soviet sources are presented in Table A1.

(iii) Finally, the composite price index for deflating material purchases by industry is derived by applying a weight of 0.68 to the price index in industry itself, of 0.21 to the price index in Agriculture and of 0.11 to the price index in Transport and Communications. This composite price index is presented in Table A2.

that "... such omissions [of purchases by industry from Agriculture and Transportation] turn out to be empirically irrelevant for [Soviet] industry as a whole" (pp. 691–92). Furthermore, it is possible to compute the magnitudes of extra-industry purchases as a proportion of *gross value of output* in industry in the following manner: Applying the proportions of extra-industry purchases of 36 to 37 percent in 1959 and 1966 to the corresponding material purchases by industry of 78.609 billion and 173.998 billion rubles estimated via the method of Section A of the Appendix, one gets the values of 28.299 and 64.379 billion rubles as estimates of extra-industry purchases by industry in 1959 and 1966, respectively. Dividing these by the corresponding values of gross outputs in industry of 118.743 and 259.266 billion rubles in 1959 and 1966, again estimated in Section A of the Appendix, the percentages of extra-industry purchases by industry as proportions of *gross value of output* in industry turn out to be 24 percent in 1959 and 25 percent in 1966. These are by no means insignificant amounts.

TABLE A2—PRICE INDEX FOR DEFLATING MATERIAL PURCHASES BY INDUSTRY[a]

| (1) 1958 | (2) 1959 | (3) 1960 | (4) 1961 | (5) 1962 | (6) 1963 | (7) 1964 | (8) 1965 | (9) 1966 | (10) 1967 | (11) 1968 | (12) 1969 | (13) 1970 | (14) 1971 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 98.68 | 100.73 | 100.00 | 101.08 | 104.28 | 106.75 | 101.47 | 105.10 | 104.02 | 109.59 | 109.20 | 112.11 | 116.17 | 118.58 |

[a] The price deflator for 1958 is used for deflating material purchases during 1955–57.

TABLE A3—DATA UNDERLYING DESAI HYBRID INDEX[a,b]

| Branch (1) | Desai Gross Value-Added (2) | Desai Weights (3) | Weitzman Weights (4) |
|---|---|---|---|
| Electricity | 1.770 | 3.87 | 5.01 |
| Fuels | 5.489[c] | 11.99 | 14.73 |
| Ferrous metals | 2.784 | 6.08 | 7.47 |
| Nonferrous metals | 1.379[c] | 3.01 | 3.70 |
| Wood & paper products | 3.896 | 8.51 | 9.84 |
| Construction materials | 2.131 | 4.65 | 6.42 |
| Chemicals | 1.939 | 4.23 | 4.36 |
| Glass and ceramics | .304[c] | .66 | .82 |
| Machinebuilding and metal working | 13.207 | 28.84 | 28.89 |
| Soft goods | 6.420 | 14.02 | 10.35 |
| Processed foods | 6.473 | 14.14 | 8.41 |

*Source:* As explained in the text.

[a] The weights are applied to the branch gross output indexes available in Soviet sources, by both Weitzman and Desai.

[b] Column (2) is shown in billion rubles, and columns (3) and (4), in percent.

[c] Weitzman estimates.

The gross value-added figures in industry derived by using the double deflation procedure are converted into indexes with 1960=100 and presented in column (3) of Table 1. The indexes of material purchases with 1960=100 are presented in column (5) of the same table.

## REFERENCES

K. Arrow, "The Measurement of Real Value-Added," tech. rept. No. 60, dept. econ., Stanford Univ. 1972.

——— et al., "Capital-Labor Substitution and Economic Efficiency," *Rev. Econ. Statist.*, Aug. 1961, *43*, 225–50.

P. David, "The Deflation of Value-Added," *Rev. Econ. Statist.*, May 1962, *44*, 148–55.

———, "Measuring Real Net Output: A Proposed Index," *Rev. Econ. Statist.*, Nov. 1966, *48*, 419–25.

P. Desai, "Soviet Industrial Production: Estimates of Gross Outputs by Branches and Groups," *Oxford Inst. Econ. Statist. Bull.*, Summer 1973, *35*, 153–71.

———, "Technical Change, Factor Elasticity of Substitution and Returns to Scale in Branches of Soviet Industry in the Postwar Period," in F. L. Altmann, O. Kýn, and H. J. Wagener, eds., *On the Measurement of Factor Productivities: Theoretical Problems and Empirical Results*, Gottingen-Zurich 1975.

B. Hansen, "Effective Protection and its Measurement," dept. econ. work. pap., Univ. California, Berkeley, Sept. 1973.

J. Thornton, "Value-Added and Factor Productivity in Soviet Industry," *Amer. Econ. Rev.*, Dec. 1970, *60*, 863–71.

M. Weitzman, "Soviet Postwar Economic Growth and Capital-Labor Substitution," *Amer. Econ. Rev.*, Sept. 1970, *60*, 676–92.

Tsentral'noe Statisticheskoe Upravlenie pri Sovete Ministrov S.S.S.R., *Narodnoe Khoziaistvo S.S.S.R.*, Moscow 1956–1971.

———, *Narodnoe Khoziaistvo S.S.S.R.*, Moscow 1922–72, Yubileinyi Statisticheskii Ezhegodnik, Moscow 1972.

———, *Promyshlennost' S.S.S.R.*, Moscow 1964.

# Public Safety

By SERGE-CHRISTOPHE KOLM*

All measures taken by the public sector have some influence upon the occurence of human deaths and other bodily injuries. And to diminish these risks is the essential or unique aim of a number of them, which include both expenditures and regulations. What is the optimum for these measures? Is the present situation close to it?

A very strong negative answer to the latter question has often been suggested by the observation of the marginal costs incurred to save lives in various sectors by various measures (or, equivalently, of the marginal efficiency of life saving expenditures). The information is plentiful for several countries; in English, let me just mention the studies by William Gorham and Dan Usher and the references they give. The result is always of the same order of magnitude: according to sectors and measures, these marginal costs or efficiencies can be multiplied by factors up to thousands. Can we conclude that human life, the most important commodity, the only one which is an indispensable complement to the enjoyment of any other, is also the one produced with the grossest misallocation and waste?

Several arguments have been and could be advanced to justify discrepancies in these marginal costs and efficiencies. Their general presentation and discussion can be found in my paper (1975). One of them is by far more important than the others, and I focus on it here. It will give results such that, most of the time, the number of dollars spent to avoid a marginal casualty must be higher the more "collective," or "public," the danger is in some sense. And, it will turn out that computations from purely technical observable data can tell us whether safety has received too little or too much emphasis in any given sector.

* School for Higher Studies in the Social Sciences, and CEPREMAP, Paris. The present English (and abbreviated) version was presented at a conference on Public Economics organized by Israel's Scientific Research Council at Kibbutz Kiryat Anavim, June 1975.

## I. Assumptions

Unavoidably, a number of simplifications must be accepted. In some cases, my model and results will be directly applicable. In others, the model must be completed by other considerations, and the role of this study then is to show the way toward important phenomena and conclusions. The main simplifications are as follows. I shall explicitly consider only death, but the same analysis applies to other bodily injuries; more generally, it is the theory of optimum allocation of nontransferable consumptions produced by both private and public inputs (education and recreation are other cases). There is only one period, short enough so that everyone's probability of dying during this period is small. The main problem considered will be to allocate public safety funds between several uses, and the only valuable effect is to reduce casualties; I assume no effective nonnegativity constraints on these different amounts. Finally, there will be no externalities, an assumption which has four main applications here. 1) There are no Pareto-irrelevant (distributional) price externalities, which means that either the considered variations affect too small a sector of the whole economy to influence prices, or that price effects are compensated. 2) There is no externality in consumption: individuals are not directly affected by the consumption-work choices of others, be it through their tastes or their choices' nonbudget constraints. 3) As a special case of 2), individuals are not so affected by the probability of death of others. 4) The death probability of individuals is not affected by the others' choices, i.e., we exclude manslaughter. (Accidents as "congestion externalities" are analyzed in my book (1970).)

The last three cases require an explanation. For individuals' choices about consumption, work, safety, etc., for both the decision and the evaluation, the biological "individual" is not, in fact, the right social

unit. To consider "the family" as the unit would be far better. This term is, however, ambiguous. Every biological individual is the center of concentric circles of individuals who are more or less directly (i.e., not through markets) concerned by, and participating in, "his" choices: himself, the nuclear family, relatives, friends, acquaintances, etc. They, more or less, have a feeling for the individual's situation and influence the corresponding choices. Then, the absence of inter-individual externality means that, in some sense, their influence is "proportional to" their feeling. This participation in choice is through suasion, "reciprocity" exchange, vicarious choice, an individual's desire to please another, etc. This paper is not the place to develop this analysis. Let me just mention its application to the question of risks of death. Who is concerned by an individual's survival probability? First, himself. Then his closest family, first for affective reasons, often with the same intensity as the individual himself, and for material and economic reasons. Then his friends, relatives, etc. Then members of the social communities to which he belongs (neighborhood, church, region, nation, mankind), with a spectrum of degrees of intensity. We are assuming here that the influences on the individual's choices which affect his death probability (as well as on other aspects of his well being) internalize the corresponding feelings, especially at the intrafamily level where they would be impossible to ignore.

## II. Optimality Conditions

Turning now to traditional economic concepts, let $i$ be an individual, $p^i$ his death probability, $v^i$ his marginal willingness to pay for $p^i$, $s^i$ the social marginal value of his wealth. We assume "normal" individuals and society so that $v^i<0$, $s^i>0$, and all the $s^i$ are equal if "distribution is optimal." The term $y^i$ is the set of parameters describing $i$'s "situation" (commodities consumed, work, location, etc.); $x_j$ is the amount of public money spent to reduce risk of type $j$; and $x$ is the vector of $x_j$. From the assumptions, $p^i$ is a function of $x$ and $y^i$ only: $p^i=p^i(x, y^i)$, where $p^i_j=\partial p^i/\partial x_j$ is the "direct" marginal effect of $x_j$ on $p^i$; we shall normally have $p^i_j<0$. The individual $i$ chooses $y^i$ in a possibility set $Y^i$ (which includes in particular its budget constraint but does not directly depend upon $p^{i'}$, and $y^{i'}$ for $i'\neq i$), so as to maximize his utility function $u^i[y^i, p^i(x, y^i)]$. It is shown in the Appendix that an optimality condition for $x_j$ is that $\sum s^i v^i p^i_j$ is the same for all $j$. This condition holds if $\sum x_j$ is given, but it also holds if we explicitly consider how this sum is obtained, for instance through taxation of any kind, as long as the $x_j$ enter in the problem's constraints only by their sum. The result can be applied to any set of risks $j$ (containing at least two $j$'s).

## III. Direct and Indirect Effects

With given functions $u^i$ and $Y^i$, the $y^i$ chosen by individual $i$, $y^{i*}$, depends upon $x: y^{i*}=y^{i*}(x)$. We assume that this function is differentiable for the present exposition, although it is not necessarily so in this field (decisions to incur or shun some kind of risk). The sensitivity (total derivative) of $p^i$ for $x_j$ is

$$P^i_j = \frac{dp^i}{dx_j} = p^i_j + \sum_k \frac{\partial p^i}{\partial y^{i*}} \frac{\partial y^{i*}_k}{\partial x_j}$$

where the $y^i_k$ are the components of $y^i$. In the normal case $0>P^i_j>p^i_j$, meaning that the increment in income received by the gift in safety through an increase in $x_j$ is redistributed among several items of consumption, including some left in safety.

It is usually possible to distinguish for each $x_j$ a subset of $y^i_k$, say those with $k\in K_j$, which are parameters chosen by the individuals and having some influence on their risk which is affected by $x_j$; thus, $\partial p^i_j/\partial y^i_k=0$ for $k\notin K_j$; we then call

$$\pi^i_j = p^i_j + \sum_{k\in K_j} \frac{\partial p^i}{\partial y^{i*}_k} \frac{\partial y^{i*}_k}{\partial x_j}$$

The normal case seems to be such that $0\geq\pi^i_j\geq p^i_j$. The case $\pi^i_j=p^i_j$ generally implies that the individual has no influence on the risk affected by $x_j$. Ordinary behavior seems also to suggest that $\pi^i_j=P^i_j$ is not infrequent.

We now have three candidates for mea-

suring the marginal efficiency of $x_j$ in number of lives saved: $\sum_i |P_j^i|$, $\sum_i |\pi_j^i|$, and $\sum_i |p_j^i|$. The "right" one may seem to be the total marginal number of lives saved, $\sum_i |P_j^i|$. However, the one which is given by the above mentioned statistics is the number of lives saved in the considered activity, $\sum_i |\pi_j^i|$. One may consider the individuals' adjustments as their own responsibility, implying that the marginal efficiency which should count for allocating $x_j$ is the marginal number of lives "directly" saved, $\sum_i |p_j^i|$. According to which number of deaths we would like to minimize, it is the equality for all $j$ of one of these three numbers which are the corresponding first-order conditions. None is, however, a priori the true optimality condition.

### IV. When Must We Save As Many Lives As Possible?

There are, however, nonfortuitous cases when $\sum_i p_j^i = \sum_i p_k^i$ is the condition for optimal allocation between $x_j$ and $x_k$. One is when $x_j$ and $x_k$ are such that, necessarily, $p_j^i$ is the same for all $i$ and $p_k^i$ is the same for all $i$. We call the corresponding risks "pure public risks." Practically, this happens for risks which are common to the whole population under consideration. The case also practically requires that individuals cannot modify this risk for themselves, and thus that $p_j^i = \pi_j^i$ and $p_k^i = \pi_k^i$. In such cases, if statistics show $|\pi_j^i| \neq |\pi_k^i|$, or if engineering computation concludes that $|p_j^i| \neq |p_k^i|$ in the actual situation, there is misallocation and funds must generally be transferred from the less to the more efficient use.

Another case, at the extreme opposite, is when $x_j$ and $x_k$ actually affect the safety of only one and the same individual. That is, $p_j^{i'} = p_k^{i'} = 0$ for all $i' \neq i$. The corresponding risks are "purely private risks" for individual $i$. We then note that the condition $p_j^i = p_k^i$ does not depend upon the fact that individual $i$ may be more able to shun one risk than the other one by his own choices, i.e., does not depend upon the differences in individual responsibility in risks.

These two cases are subcases of the one where $x_j$ and $x_k$ affect in a purely public way the safety of the individuals of some subsociety, and do not affect the others. That is, there is a set $I$ of $i$ such that $p_j^i = p_k^i = 0$ for $i \notin I$, and $p_j^i$ are the same for all $i \in I$ and $p_k^i$ are the same for all $i \in I$.

A third case would be when all $V^i = v^i s^{i}$'s are equal: then $\sum_i p_j^i$ must be the same for all $j$. But this certainly does not hold. Even apart from inequalities in $s^i$, the $v^i$ are far from equal, due to differences in individuals' situations and tastes (some people never take a plane, others play Russian roulette).

In all these cases but the last, the result does not depend upon the $V^i$, and therefore not on the $v^i$ or $s^i$. Thus it is independent of the income distribution, of individual tastes, and of costs of raising public funds. The most significant of these cases is that of pure public risks: the allocation of safety funds among risks must save as many lives as possible.

### V. The Technological Optimality Condition

With $s^i v^i = V^i$, the optimality conditions are $\sum_i V^i p_j^i = \lambda$. We write $p_i$ for optimum $p_j^i$ when $j$ is a pure private risk for $i$. The condition gives $V^i p_i = \lambda$. Then, for any general risk $j$, the condition is $\sum_i p_j^i / p_i = 1$. The interest of this formula is that $p_j^i/p_i$ is the rate of substitution of purely private safety costs to $x_j$ in $p^i$. In particular, it may be a purely private safety expenditure in the same field where $x_j$ intervenes. Then, $p_j^i/p_i$ are often computable through engineering data or from casualty statistics. We would generally conclude that $x_j$ is too small or too large from the fact that $\sum_i p_j^i/p_i$ is larger or smaller than one.

### VI. Pure Public and Private Risks

Let us now call $p_0$ the optimum $p_j^i$ for any pure public risk $j$. The above optimality condition gives $1/p_0 = \sum 1/p_i$. Or, $n$ being the number of persons in the society,

$$\frac{1}{np_0} = \frac{1}{n} \sum \frac{1}{p_i}$$

$n|p_0|$ is the direct marginal efficiency of purely public safety in number of lives saved. It must thus be the harmonic mean of the like efficiencies of purely private safety, $|p_i|$. Thus, if $m = Min_i |p_i|$ and $M =$

$Max_i\,|p_i|$, $m \leq n|p_0| \leq M$ must hold. Furthermore, from an easy application of Cauchy's inequality, one must have $n\,|p_0| \leq (1/n)\sum |p_i|$, that is to say, one extra dollar must save more lives if it is equally shared between purely private safeties than if it is used for pure public safety; both allocations are "egalitarian" in some sense. It may seem surprising that equal sharing, which is arbitrary from efficiency viewpoints, is nevertheless the most efficient way of saving lives. Schweitzer's and Sisha-Mond's inequalities furthermore show that

$$\frac{1}{n}\sum |p^i| / n |p_0| \leq (M+m)^2/4Mm$$

and

$$\frac{1}{n}\sum |p^i| - n|p_0| \leq (M^{1/2} - m^{1/2})^2$$

The equalities hold everywhere if and only if all $p_i$ were equal, which is unlikely since it would also imply the equality of all $V^i$ at the optimum. Thus, if a nonpure public risk $j$ is close enough to a pure public risk in the sense that the dispersion of the $p_j^i$ is low enough, the same inequalities must hold when $\sum_i |p_j|$ replace $n|p_0|$. In particular, $\sum_i |p_j^i| \leq (1/n)\sum |p_i|$. It is, however, the comparison between general $\sum_j |p_j^i|$ and $n|p_0|$ which is the most interesting one: this is the topic of Section VIII.

## VII. Public or Private Provision

Before leaving purely private risks, we must note a difference between two kinds of safety expenditures according to who decides about them: the public sector or the individual. There are publicly provided pure private safety and privately provided pure private safety. We considered the former. For the latter, the individual's choice does not necessarily conform to the social optimum: if we call $p_i'$ the derivative of $p^i$ for any safety expenditure chosen by individual $i$, this choice is such that $v^i p_i' = 1$. This private choice would coincide with what would be the optimum one if the public sector were in charge of this decision if $s^i = \lambda$, giving $p_i' = p_i$ at optimum. This would happen if lump sum transfers between individual $i$'s wealth and the public safety fund $\sum x_j$ were possible. Otherwise, generally $p_i' \neq p_i$. More precisely, at optimum $p_i'/p_i = s^i/\lambda$, which means that the marginal rate of substitution of publicly provided to privately provided purely private safety funds for $i$ must be equal to the social marginal value of $i$'s wealth measured in public safety dollars. Now, statistics sometimes provide $p_i'$ more easily than $p_i$; for instance, in the case of road traffic safety, accident statistics based on cars (a private choice) give $p_i'$, whereas statistics based on the condition of roads (a public safety choice) give $p_j^i$. If "income distribution is optimal," $s^i$ is the same for all $i$ by definition, and the first-order optimality condition for the allocation of the $x_j$ is that $\sum_i p_j^i/p_i'$ is the same for all $j$. If, furthermore, there can be lump sum transfers between individuals' incomes and the public safety fund, these conditions become that $\sum_i p_j^i/p_i' = 1$ for all $j$. And the actual discrepancies between these values indicate how the $x_j$ must be reallocated.

## VIII. General and Pure Public Risks

Optimum allocation between purely public and general safety imply

$$p_0 = \sum_i \frac{V^i}{\sum V^i} p_j^i = \sum_i \frac{1/p_i}{\sum 1/p_i} p_j^i$$

and the marginal number of lives saved by $x_0$ is

$$n|p_0| = \sum_i \frac{V^i}{\overline{V}} |p_j^i|$$

where $\overline{V} = \sum V^i/n$ is the average $V^i$. Therefore $n|p_0| < \sum_i |p_j^i|$ if the $p_i$ and the $p_j^i$ are similarly classified by $i$, and the reverse inequality holds if they are inversely classified by $i$ (i.e., $(p_i - p_{i'})(p_j^i - p_j^{i'})$ for all pairs $i, i'$ is positive in the former case and negative in the latter one).

Let $\pi^{ij}$ be the probability that the cause of risk $j$ kills individual $i$. It is an additive part of $p^i$. It depends upon $x_j$ the amount of public money which diminishes this risk for all $i$, and upon $y_j^i$ the sum spent to diminish this risk for individual $i$ only, $x_j$ and $y_j^i$ influence $p^i$ through $\pi^{ij}$ only. We also assume that $\pi^{ij} = \pi^j(x_j, y_j^i)$; i.e., the only reason why

$\pi^{ij}$ depends upon $i$ is that the $y_j^i$ differ. This is a frequent situation which only means that the effect of $x_j$ and $y_j^i$ on $\pi^{ij}$ is a purely technical matter, general and not dependent upon other individual characteristics; $\pi^j$ is a decreasing function of both its arguments. But, being a probability, it is constrained by $\pi^{ij} \geq 0$. It is small by assumption, and in ordinary cases very small. We consider the situation at the optimum. We certainly have $\partial^2 \pi^j / \partial y_j^{i2} > 0$ because of $\pi^j$'s positivity and smallness; this indicates a point of decreasing returns to scale of $y_j^i$ expenditure on safety. But $\partial \pi^j / \partial y_j^i = p^i$ and $\partial \pi^j / \partial x_j = p_j^i$. Thus $p_i$ is an increasing function of $y_j^i$, and therefore $p_i$, $y_j^i$, $-\pi^{ij}$ are similarly ordered by $i$. Also $p_j^i$ is similarly ordered or, on the contrary, inversely ordered, by $i$, depending on whether it is an increasing or a decreasing function of $y_j^i$, i.e., on whether $\partial^2 \pi^j / \partial x_j \partial y_j^i$ is positive or negative, i.e., on whether $x_j$ and $y_j^i$ are substitutes or complements in providing safety.

To conclude, *a public expenditure on a general public safety must marginally save more or fewer lives than one on a pure public safety depending on whether it is a substitute or a complement to private safety measures* (i.e., the marginal cost of saving a life must be lower or higher in this manner than from pure public risks depending on this condition).

In most cases, public and private safety measures happen to be substitutes rather than complements. This is for instance the case for most road (public) and car (private) safety improvements, for firemen and private fire protection (extinguishers, use of fireproof material, exits), for police and private watchmen or systems preventing entry, for most home and hospital cures, for vaccination and public sanitation, for public air pollution prevention and private air filters, etc. There are however cases when they are complements: for instance, various aspects of health care, or car lights and road signals, or, generally, private alarm systems (things or men) and the public service they call (fire, police, accident rescue).

The most common case thus is $n|p_0| < \sum_i |p_j^i|$ at the optimum, i.e., an extra dollar must save fewer lives from a pure public risk than from nonpure public risks, or, we must pay more to save an extra life from a pure public risk than elsewhere. The reason for it as shown above is that the nonpure public safety measure saves on the average "cheaper" lives than the purely public one because the more valuable ones are more sheltered from this risk by specific protection—a discrimination which is not possible for purely public risks.

The difference in death's marginal costs may have to be substantial. If risk averters (i.e., high $|v^i|$ people) completely shun risk $j$ so that $p_j^i = 0$ for these $i$, whereas other people care little (low $|v^i|$), then $\sum_i (V^i/\overline{V})|p_j^i| = n|p_0|$ may be much smaller than $\sum_i |p_j^i|$.

## IX. General Risks

Now calling $\rho_j$ the correlation coefficient between $V^i$ and $p_j^i$, we have

$$\sum_i V^i p_j^i = \overline{V} \cdot (1 + \rho_j) \cdot \sum_i p_j^i$$

which shows that the marginal cost of saving a life by $x_j$ must be proportional to $1 + \rho_j$ ($\rho_j = 0$ for a pure public risk and the above argument is that $\rho_j < 0$ for most of the other cases). It is also possible to distinguish the factors in $V^i = s^i v^i$ and analyze the effect on the optimum marginal costs of life of the correlations between $s^i$ (income distribution), $v^i$ (individual risk aversion), and $p_j^i$.

## X. Conclusion

Given its importance, the economic value of human life has received little attention. It is hoped that this paper sheds some light on some of these important aspects, in particular on the question of the inequality of values for different sectors and decisions, especially public decisions. Among the results are the following: There generally must be inequalities. An exception is the case of purely public risks where as many lives as possible must be saved by the appropriate allocation of safety funds. For other safety measures, life must generally be cheaper or more expensive than for purely public safety

depending on whether this expenditure is a substitute or a complement to purely private safety; most often, it must be cheaper. Finally, it has been shown how statistical or engineering data may indicate whether safety is too low or too high in any given sector.

## Appendix

The usual notations are used for the derivatives for vectors and of vectors. We call $v^i_y$ the vector of individual $i$'s willingness to pay for the components of $y^i$ when $p^i$ is given. We note $p^i_y$ for $p^i_{y^i}$. Given $x$, individual $i$ chooses an $y^i$, $y^{i*}(x) \in Y^i$. Either $v^i p_y + v^i_y = 0$, or $y^{i*}(x)$ is on the border of $Y^i$ and the vector $v^i p_y + v^i_y$ for $y^i = y^{i*}(x)$ is within the cone of normals to $Y^i$ at this point. Thus, when $x$ differentially varies, we have

$$(v^i p^i_y + v^i_y)\Big|_{y^i = y^{i*}(x)} \cdot y^{i*}_x(x) \cdot dx = 0$$

Therefore, individual $i$'s marginal willingness to pay for a change $dx$ in $x$ is the value at $y^i = y^{i*}(x)$ of

$$[v^i p^i_x + (v^i p^i_y + v^i_y) y^{i*}_x] \cdot dx = v^i p^i_x \cdot dx$$

whence the mentioned optimality condition. The same result can be shown to hold when $y^{i*}(x)$ is not differentiable, a not unimportant case (choices to incur or not some risk—for instance, to take a plane or not).

## REFERENCES

C. Abraham, Goldberg, and J. Thédie, "Le choix des investissements routiers," *Transports*, Nov. 1960.

C. Abraham and J. Thédie, "Le prix d'une vie humaine dans les décisions économiques," *Revue Française de Recherche Opérationnelle*, 3ème trimestre 1960.

J. Drèze, "L'utilité sociale d'une vie humaine," *Revue Française de Recherche Opérationnelle*, 2ème trimestre 1962.

G. Fromm, "Comment" on Schelling's paper, in S. B. Chase, Jr., ed., *Problems in Public Expenditure Analysis*, Washington 1968.

W. Gorham, "Some Use of Quantitative Analysis to Improve the Allocation of Public Funds," in J. Paelinck, ed., *Programming for Europe's Collective Needs*, Amsterdam 1970.

J. H. Jones, *Road Accidents—Report submitted to Minister of Transport*, London 1946.

M. Jones-Lee, "Valuation of Reduction in Probability of Death by Road Accident," *J. Transport Econ. Policy*, Jan. 1969.

S. Ch. Kolm, "Concernements et décisions collectifs, une contribution à l'analyse économique des sociétés," *Analyse et Prévision*, 1967.

———, *Le Service des Masses*, Paris 1970.

———, "Connaissance des coûts et valeurs d'environnement," in *Estimation des dommages d'environnement*, O.C.D.E., 1972–74; English translation (O.E.C.D.).

———, "L'inégalité des valeurs des vies humaines," *Cahiers du Séminaire d'Econométrie*, 1975.

R. McKean, "Comment" on Schelling's paper, in S. B. Chase, Jr., ed., *Problems in Public Expenditure Analysis*, Washington 1968.

E. J. Mishan, "Evaluation of Life and Limb: A Theoretical Approach," *J. Polit. Econ.*, July-Aug. 1971, *79*, 687–705.

G. Morlat et al., "Recherches sur les coûts de protection de la vie humaine," Centre de Recherche Opérationnelle, Institut de Statistique des Universités de Paris.

D. J. Reynolds, "The Cost of Road Accidents," *J. Royal Statist. Soc.*, 1956, pt. 4, *119*, 393–408.

T. Schelling, "The Life you Save May be Your Own," in S. B. Chase, Jr., ed., *Problems in Public Expenditure Analysis*, Washington 1968.

J. Thédie, "Valeur économique de la sécurité routière," *La Route*, 1958.

D. Usher, "An Imputation to the Measure of Economic Growth for Changes in Life Expectancy," in M. Moss, ed., *The Measurement of Economic and Social Performance*, Nat. Bur. Econ. Res. *Stud. in Income and Wealth*, Vol. 38, New York 1974.

CERMAP, "Le dépistage précoce du cancer du col de l'utérus. Etude économique," juillet 1968.

# A State-Preference Approach to the Precautionary Demand for Money

*By* M. L. CROPPER*

Much attention has been devoted in the literature to the transactions demand for money and the asset demand for money. The precautionary motive for holding cash balances, however, is seldom analyzed. A notable exception is S. C. Tsiang's inventory-theoretic analysis, which integrates precautionary demand into an inventory model of the transactions demand for cash. Precautionary demand can, however, be approached as a problem in portfolio theory using a state-preference approach, as John Hicks (1967) once suggested.[1] The states of nature in this case represent the times at which an investor may, for reasons beyond his control, have to liquidate his portfolio. It is this uncertainty regarding future cash requirements, combined with the fact that some assets are neither readily nor costlessly convertible into cash, which gives rise to the precautionary demand for money.

Consider the situation of an investor who is allocating his wealth among assets of varying degrees of liquidity to maximize expected utility of wealth at the end of his horizon. The investor believes that certain events (outside of his control) may occur which will force him to liquidate his portfolio before the end of his horizon. If liquidation is not called for, the portfolio will be held until the end of the horizon and then sold. Whether or not liquidation will actually be required is not known until the time the cash is needed. Thus, if the portfolio must be sold before the end of the period, it must be sold on short notice. This means that the investor will be able to sell his less liquid assets only at a discount, the size of the discount varying inversely with the length of notice given before the asset is sold. The individual thus has an incentive to hold liquid assets, say, cash, to reduce losses which may possibly be incurred if he must suddenly liquidate his portfolio. Money held for this purpose may be labelled precautionary cash balances. An increase in the probability that the investor will have to liquidate his portfolio in the near future should increase this demand.

The remainder of this paper investigates the precautionary demand for money in the framework outlined above. Using two concepts of liquidity, I derive conditions under which an increase in the probability of having to liquidate one's portfolio on short notice will increase the precautionary demand for money; that is, conditions are obtained under which a well-behaved precautionary demand for money, in the above sense, can be said to exist.

In Section I the concept of liquidity is discussed at some length. This discussion is necessary before liquidity can be formally incorporated into a portfolio selection model. The formal model is presented in Section II.

## I

A liquid asset according to J. M. Keynes' definition is one which is " . . . more certainly realizable on short notice without loss" (p. 67). This is usually interpreted to mean that *il*liquid assets are either not realiz-

* Assistant professor of economics, University of California, Riverside. I gratefully acknowledge helpful comments from two anonymous referees.

[1] I have not however followed the model outlined by Hicks (1967, pp. 31–34). In that model, an illiquid asset is one which incurs high transactions cost, e.g., brokerage fees, whenever it is traded, *regardless* of the length of notice given before the transaction takes place. (This use of the term liquidity deviates somewhat from that followed by Keynes and Tsiang.) The longer the investor's horizon, the more incentive he will have to invest in illiquid assets since transactions costs will decline as a percentage of asset yield as the holding period is lengthened. Put somewhat differently, Hicks is saying that it will pay to invest in illiquid assets only if they can be held long enough to compensate for their high transactions costs. Precautionary demand enters when the planned date of realization is uncertain.

able with certainty or that they can be converted into cash on short notice, only at a loss. The size of the loss incurred is assumed to vary inversely with the length of notice given before the asset is sold. This is the common interpretation of the term liquidity (see Hicks, 1962; Tsiang); however, a little reflection indicates that this interpretation is not meaningful in all market situations. In a perfectly competitive capital market, for example, there is no obvious reason why the price at which an asset is sold should vary directly with the length of notice given before the asset is sold. Therefore, before liquidity can formally be incorporated into a portfolio selection model, some further discussion is clearly required. To give the term more meaning, I shall discuss two sources of uncertainty in capital markets.

Uncertainty in most portfolio selection models is uncertainty about which state of nature will occur at some future time t.[2] Investors are assumed to know security prices contingent on the state of the world at that time. Because an investor is uncertain exactly what the state of the world will be, he has a subjective probability distribution over asset prices. Once time t has arrived, however, the state of the world is known and there is a single known asset price. (Capital markets are assumed to be perfect in all respects other than knowledge of the future.)

In the above model the individual believes that the price of each security at time t is determined by forces external to him, for example, political conditions in the United States, the world demand for steel, etc. Once the outcome of these forces is known a unique price is determined for each security. The investor has no reason to believe that this price will be altered should he have to sell the security at time t. Thus there is no reason to believe that the investor's distribution over future asset price should be affected by the length of notice given before an asset is sold. By this reasoning, the "realizable on short notice without loss" aspect of the term liquidity really has no meaning in a portfolio model in which uncertainty is only of a state-of-nature variety. One asset can be more liquid than another in this context only if it is "more certainly realizable" regardless of the length of notice given before it is sold. Thus, in standard portfolio selection models, "more liquid" must be synonymous with "less risky," and a perfectly liquid asset must necessarily be a riskless asset.

Keynes' definition of liquidity is meaningful, however, in markets in which investors are imperfectly informed about opportunities to buy and sell securities. In markets in which information is imperfect there will usually be a distribution of prices for any security at any time, rather than a single equilibrium price.[3] In such a market an investor has a subjective distribution over asset yield (price) because he is uncertain about the actual price at which he will sell a security. The moments of the subjective distribution will generally depend on the amount of price information the individual can obtain by sampling bids of prospective buyers.

Assuming that sampling takes time, one may argue that an investor who does not have time to search out potential buyers, for example, because he must sell his assets on short notice, will most likely have to sell his assets at a lower price than an investor who has time to acquire a larger sample of bids. Reasoning in this way, one may interpret "realizable on short notice without loss" to mean that illiquid assets can be sold on short notice only at a discount. This discount arises because the individual does not have time to obtain more price information, and it should therefore vary inversely with the length of notice given before the asset is sold. This aspect of liquidity is usually re-

[2] The uncertainty in Markowitz-Tobin portfolio selection models is implicitly of the state-of-nature variety. In these models investors have probability distributions over security prices at some time in the future, which are by assumption independent of current prices. These distributions may be regarded as having been derived from a knowledge of state-contingent returns and a probability distribution over states when the number of states is infinite. It should be emphasized that state-of-nature uncertainty can exist even when capital markets are perfect, in the sense that each asset has a single known equilibrium price.

[3] The uncertainty in recent job-search models is of this variety.

ferred to as marketability.

In the model below we are interested not in the price at which an illiquid (imperfectly marketable) asset is actually sold, but in the effect that having suddenly to liquidate the asset will have on the investor's subjective distribution over asset yield. That is, we wish to know how the belief that an illiquid asset must be sold on short notice will alter the investor's subjective probability distribution over yield. If an investor believes that by increasing the size of his sample he will reduce the variability of the distribution and increase its mean, it follows that the subjective distribution should have greater variability (a more precise definition of greater variability is given below) and a lower mean if the investor must sell the security on short notice than if he has more time to sample potential buyers.[4] With this interpretation of liquidity in mind, we turn to the formal model.

## II

Assume that an investor must divide his initial wealth $W_0$ between two securities: money, a perfectly liquid asset paying zero return with certainty, and an illiquid asset, the return on which is a random variable. One of two possible events which may be thought of as states of nature must occur before the end of the investor's horizon. The first event, which we shall call state 1, is the occurrence of some emergency which requires immediate cash payment. If state 1 occurs the investor must sell his portfolio at once. State 2 is the absence of such an emergency. If state 2 occurs the portfolio will be held to the end of the horizon and then sold.

Money, because it is a perfectly liquid asset, pays a certain return (of zero) in both states.[5] The return on the illiquid asset is $r_1$ in state 1 and $r_2$ in state 2. It is assumed that this asset is illiquid because it is imperfectly marketable. Thus the distribution over $r_1$ should be more variable and should have a lower mean than the distribution over $r_2$, since the investor will not have as much time to sample price information in state 1 as he will if state 2 occurs.[6] It is further assumed that $r_1$ and $r_2$ must lie in the interval $[-1, \infty)$. This implies that the investor's wealth can never fall below 0.[7]

The investor's objective is to set $a$, the fraction of $W_0$ invested in the illiquid asset, to maximize expected utility of wealth at the time his portfolio is liquidated. Thus the investor seeks to

$$\max_a \{pE[U(W_1)] + (1-p)E[U(W_2)]\},$$

$$W_1 = W_0(1 + ar_1) \tag{1}$$
$$W_2 = W_0(1 + ar_2)$$

where $p$ is the subjective probability that state 1 will occur and $(1-p)$ is the subjective probability attached to state 2. The investor's utility function is assumed to be identical in both states and to obey

$$U' > 0 \quad \text{and} \quad U'' < 0 \tag{2}$$

Because we are interested only in a solution for which $0<a<1$, I shall make the following assumptions which guarantee an interior solution to (1).

$$U'(0) = \infty \tag{3}$$

$$E(r_1) > 0, \qquad E(r_2) > 0 \tag{4}$$

First-order and second-order conditions for a maximum are, respectively,

$$pW_0E[r_1U'(W_1(a))] + (1-p)W_0E[r_2U'(W_2(a))] = 0 \tag{5}$$

[4] This is a tricky point. The investor is making a judgment in the present, before any sample is taken, about what the shape of his subjective distribution will be in the future, after the sample has been taken. The belief before the sample is taken that the variability of the distribution can be reduced by taking a larger sample corresponds to what in Bayesian terminology is referred to as preposterior analysis.

[5] The analysis may be extended with little difficulty to the case where money pays a nonzero return with certainty by reinterpreting $r$ as the yield differential between liquid and illiquid assets. As this does not significantly affect the results, I present only the case where the return on money is zero.

[6] Any other sources of uncertainty, e.g., state-of-nature uncertainty, by assumption do not differ between the two states.

[7] It also implies that the limits of integration in equations (6) and (7) below are independent of $r$; hence, differentiation is permissible within the integral sign.

$$\frac{da}{dp} = \frac{-W_0\{E[r_1U'(W_1(a))] - E[r_2U'(W_2(a))]\}}{pW_0^2E[r_1^2U''(W_1(a))] + (1-p)W_0^2E[r_2^2U''(W_2(a))]} \tag{7}$$

$$pW_0^2E[r_1^2U''(W_1(a))] + (1-p)W_0^2E[r_2^2U''(W_2(a))] < 0 \tag{6}$$

We now wish to see whether an increase in $p$, the probability that the investor must suddenly liquidate his portfolio, will increase the demand for money. This is, of course, equivalent to determining whether $da/dp$ is in fact negative. Applying the implicit function theorem to (5), an implicit function of $a$ and $p$, yields equation (7), shown above.

The denominator of (7) must be negative if the second-order condition for a maximum, equation (6), is satisfied. Thus $da/dp \gtreqless 0$ as $E[r_2U'(W_2)] \gtreqless E[r_1U'(W_1)]$. We can make statements about the sign of the numerator of (7) under alternative assumptions about the distribution functions $F_1(r_1)$ and $F_2(r_2)$. For reasons elaborated in the previous section, I argue that $F_2(r_2)$, the distribution function over the yield on the illiquid asset in state 2, should have higher mean and lower variability than $F_1(r_1)$.

We consider first the case where $r_2 = r_1 + k$, $k > 0$, so that for every possible value of $r_1$ there is a corresponding value of $r_2$ which is greater by $k$ and has the same probability occurrence. This implies that $F_1(r_1)$ and $F_2(r_2)$ are identical distributions, but that the mean of $F_2(r_2)$ is greater than the mean of $F_1(r_1)$. In this case the sign of $da/dp$ depends on whether the gain in utility from investing an additional dollar in the illiquid asset, $rU'[W_0(1+ar)]$, is increasing or decreasing in $r$.

From the assumption that $r_2 = r_1 + k$ it follows that

$$\int r_1U'[W_0(1+ar_1)]dF_1(r_1) \gtrless \int r_2U'[W_0(1+ar_2)]dF_2(r_2) \tag{8}$$

as

$$\frac{\partial}{\partial r}\{rU'[W_0(1+ar)]\} \lessgtr 0 \tag{9}$$

From equation (7) we know that the sign of the inequality in (8) determines whether the fraction of initial wealth invested in the illiquid asset increases or decreases as the probability of having suddenly to liquidate the portfolio increases. Thus, assuming that the illiquid asset has lower expected return if the portfolio must be sold on short notice, the proportion of $W_0$ invested in this asset will increase if $\partial/\partial r\{rU'(W)\} < 0$ and will decrease if $\partial/\partial r\{rU'(W)\} > 0$. That is, an increase in the probability of having suddenly to liquidate the portfolio will cause an investor to put a smaller proportion of his wealth in the less liquid asset and will increase his precautionary demand for money, provided

$$\frac{\partial}{\partial r}\{rU'[W_0(1+ar)]\} > 0 \tag{10}$$

It is easily shown that (10) is equivalent to $U' + W_0arU'' > 0$. This condition can more meaningfully be stated in terms of the relative risk-aversion index $R = -WU''/U'$. Thus

$$\frac{\partial}{\partial r}\{rU'[W_0(1+ar)]\} = U'\left(1 - \left[\frac{ar}{1+ar}\cdot R\right]\right) \tag{11}$$

A necessary and sufficient condition for the right-hand side of (11) to be positive is that $R < (1+ar)/ar$. I argue that this is in fact likely to be the case. Since $a$ is a fraction and $r$, the single-period rate of return on the risky asset, is presumably also a fraction, it is very likely for $r > 0$ that $(1+ar)/ar > 2$. Thus a sufficient condition for an increase in $p$ to reduce the demand for the illiquid asset and to increase the demand for precautionary cash balances is that $R$, the relative risk-aversion index, be less than two. The literature on the theory of risk aversion suggests that this condition is indeed likely to be

satisfied.[8] Thus for most utility functions an increase in the probability of having to liquidate one's portfolio on short notice should increase the precautionary demand for money if liquidity is defined in terms of differences in the means of $F_1(r_1)$ and $F_2(r_2)$.

Let us now consider a different interpretation of the term liquidity. Above it was assumed that only the mean of the subjective probability distribution was affected by forced sale of the portfolio. We now assume that the distribution over $r_2$ has the same mean as the distribution over $r_1$, but less variability. This is justified on the grounds that obtaining a larger sample can be expected to reduce the dispersion of the distribution over $r$.

Variability will be defined in the manner suggested by Michael Rothschild and Joseph Stiglitz (1970), viz.,

$$r_1 \underset{d}{=} r_2 + Z \tag{12}$$

where $E(Z|r_2)=0$. This should be read "$r_1$ has the same distribution as $r_2$ plus noise." Using this definition of variability we can apply the following theorem (pp. 237–38) to make statements about the sign of the numerator of equation (7).

THEOREM (Rothschild and Stiglitz): *The following statements are equivalent*:

$$Y \underset{d}{=} X + Z, \qquad E(Z|X) = 0 \tag{13}$$

$$EU(X) \geq EU(Y), \qquad U'' < 0 \tag{14}$$

As Rothschild and Stiglitz suggest (p. 67), this theorem can be applied to problems of expected utility maximization in which an agent seeks to

$$\max_{\alpha} \int U(X, \alpha)dF(X) \tag{15}$$

The optimal $\alpha$ must satisfy

$$\int \frac{\partial U(X, \alpha)dF(X)}{\partial \alpha} = EU_\alpha(X, \alpha) = 0 \tag{16}$$

If $U_\alpha$ is monotone decreasing in $\alpha$ and if $U_\alpha(X, \alpha)$ is a concave function of $X$, then an increase in riskiness in the sense of (13) will by the theorem stated above imply

$$EU_\alpha(X, \alpha) \geq EU_\alpha(Y, \alpha) \tag{17}$$

In order to apply this result in the present case, we must see under what conditions $g(r)=rU'[W_0(1+ar)]$ is concave in $r$. (That $g(r)$ is a decreasing function of $a$ is implied by equation (2).) For all utility functions for which $g(r)$ is concave, it will be true by (12) and (17) that $E[r_2U'(W_2)]\geq E[r_1U'(W_1)]$. Hence for all utility functions which imply $g(r)$ concave, $da/dp<0$, and a well-behaved precautionary demand for money can be said to exist.

A necessary and sufficient condition for $g(r)$ concave is

$$\begin{aligned} g''(r) &= aW_0(aW_0rU'''+2U'') \\ &= aW_0[(W-W_0)U'''+2U'']<0 \end{aligned} \tag{18}$$

Following Rothschild and Stiglitz, the term in brackets can be written in terms of the expressions for relative and absolute risk aversion, yielding

$$\begin{aligned} g''(r) = aW_0[U''(1 - R + W_0A) \\ + U'(W_0A' - R')] < 0 \end{aligned} \tag{19}$$

where $R=-WU''/U'$ denotes relative risk aversion and $A=-U''/U'$ denotes absolute risk aversion. Equation (19) along with previous assumptions implies that a sufficient condition for $g(r)$ concave is that the investor's utility function exhibit nonincreasing absolute risk aversion, nondecreasing relative risk aversion, and a relative risk-aversion index less than $1+W_0A$.

Arrow (ch. 3) has convincingly argued that nondecreasing relative and nonincreasing absolute risk aversion are reasonable conditions to impose on the utility function. If absolute risk aversion is decreasing, investors are willing to risk larger amounts as their wealth increases, a prediction which

[8] Arrow (ch. 3) for example argues that $R$ should be in the neighborhood of one. His argument is based on the fact that if $U(W)$ is bounded from below, then $R$ cannot approach a limit *above* one as $W$ approaches 0, while if $U(W)$ is bounded from above, $R$ cannot approach a limit *below* one as $W$ approaches infinity. From this Arrow concludes that "... broadly speaking, the relative risk aversion must hover around 1, being, if anything, somewhat less for low wealths and somewhat higher for high wealths" (p. 98).

agrees with casual observation. Nondecreasing relative risk aversion implies that the wealth elasticity of demand for cash balances is greater than or equal to one. As Arrow notes (p. 103), this prediction agrees with empirical findings by Milton Friedman, Richard Selden and others.

The condition that $R$ be less than $1+W_0A$ does not easily admit of an intuitive interpretation; however, it can be shown that this condition is satisfied by many of the utility functions in the constant relative and constant absolute risk aversion classes. Utility functions which satisfy this property include the logarithmic utility function, $U(W)=log\ W$, and all constant relative risk-aversion utility functions, $U(W)=(1-b)W^{(1-b)}$, for which $0<b=R<(1+ar)/ar$.[9] As I have argued above for $R>0$ $(1+ar)/ar$ is very likely to be $\geq 2$. Therefore all constant relative risk-aversion utility functions for which $R<2$ will most likely satisfy the requirement that $R<1+W_0A$. The constant absolute risk-aversion utility function $U(W)=-e^{-bW}$, $b>0$, satisfies the property $R<1+W_0A$ provided $b<1/arW_0$.[10] Since $a$ is a fraction and $r$ is likely to be a fraction also, the condition $b<1/(arW_0)$ does not appear unduly restrictive. These considerations suggest that for a large class of utility functions, the present framework leads to a well-defined precautionary demand for money when liquidity is used in the sense of "less variability."

Before concluding, note that the model developed above yields another implication which agrees with intuition. Assuming that the utility function exhibits nondecreasing relative risk aversion, it can be shown (see the Appendix) that the proportion of initial wealth invested in liquid assets for precautionary purposes should increase as wealth increases.

## III

The purpose of this paper is three-fold: 1) to indicate how the precautionary demand for money can be incorporated into a portfolio selection model, albeit an extremely simplified portfolio selection model, using the notion of states of nature; 2) to give a possible interpretation to the term liquidity in the context of such a model; 3) to determine precisely what restrictions must be placed on an investor's utility function for a well-defined precautionary demand for money to exist.

I have argued that in a market in which there is a distribution of prices for any security rather than a single equilibrium price, it is reasonable to assume that an investor will perceive the distribution of prices with which he is faced as being more variable if he must sell the security on short notice than if he has more time in which to sample potential buyers. If this is the case, the investor will under the assumptions made above have a positive precautionary demand for money. Furthermore, an increase in the probability that his portfolio will have to be sold on short notice will increase the proportion of the portfolio invested in liquid assets, if the investor has a utility function which exhibits nonincreasing absolute and nondecreasing relative risk aversion and has a relative risk-aversion index which is less than $1+W_0A$. It can in addition be shown that the proportion of initial wealth invested in liquid assets for precautionary purposes should increase as wealth increases, provided that the utility function exhibits nondecreasing relative risk aversion.

## Appendix

The purpose of this Appendix is to show that the proportion of initial wealth invested in liquid assets for precautionary purposes will increase as wealth increases if the investor's utility function exhibits nondecreasing relative risk aversion. To demonstrate this, it is sufficient to show that

[9] For the logarithmic utility function, $R=1$ and $A=W^{-1}$; hence $W_0A=W_0W^{-1}$. The condition $R<1+W_0A$ is always satisfied since by assumption $W$ can never become negative. For the utility function $U(W)=(1-b)W^{(1-b)}$, $1+W_0A=1+bW_0W^{-1}$. The desired property $R<1+W_0A$ is equivalent to $b<1+bW_0W^{-1}$ in view of the fact that $R=b$. This will hold only if $b<1/(1-W_0W^{-1})$, which is equivalent to $b<(1+ar)/ar$ after substitution from $W=(1+ar)W_0$.

[10] In the case of the constant absolute risk aversion utility function $U(W)=-e^{-bW}$, $R=bW$, and $A=b$. The condition $R<1+W_0A$ is thus equivalent to $bW<1+bW_0$, which implies $b<(W-W_0)^{-1}$. Using the fact that $W=(1+ar)W_0$ this condition becomes $b<(arW_0)^{-1}$.

$$\text{(A2)}\quad \frac{da}{dW_0} = -\frac{pE[r_1U'(W_1)]+(1-p)E[r_2U'(W_2)]+pE[r_1W_1U''(W_1)]+(1-p)E[r_2W_2U''(W_2)]}{pW_0^2E[r_1^2U''(W_1)]+(1-p)W_0^2E[r_2^2U''(W_2)]}$$

$$\text{(A1)}\qquad R' \geq 0 \Rightarrow \frac{da}{dW_0} \leq 0$$

Implicit differentiation of the first-order condition for a maximum (equation (5)) with respect to $a$ and $W_0$ yields equation (A2). The denominator of (A2) is negative if the second-order condition for a maximum is satisfied. The first-order condition for a maximum implies that the first two terms in the numerator of (A2) equal zero. Hence to establish (A1) it is sufficient to show that

$$\text{(A3)}\quad R' \geq 0 \Rightarrow pE[r_1W_1U''(W_1)] + (1-p)E[r_2W_2U''(W_1)] \leq 0$$

This is easily demonstrated following a method of proof due to Arrow (p. 120).

If relative risk aversion is nondecreasing then $R(W_1)\geq R(W_0)$ for $r_1>0$ and $W_1= =W_0(1+ar_1)$. By definition $R(W_1)= -W_1U''(W_1)/U'(W_1)$, hence

$$\text{(A4)}\quad r_1W_1U''(W_1) \leq -R(W_0)U'(W_1)r_1$$

This same inequality also holds for $r_1\leq 0$. Taking the expectation of (A4) with respect to $r_1$,

$$\text{(A5)}\quad E[r_1W_1U''(W_1)] \leq -R(W_0)E[U'(W_1)r_1]$$

Because a similar inequality holds for $R(W_2)$ we have

$$\text{(A6)}\quad pE[r_1W_1U''(W_1)] + (1-p)E[r_2W_2U''(W_2)] \leq -R(W_0)\{pE[U'(W_1)r_1] + (1-p)E[U'(W_2)r_2]\}$$

But the first-order condition for a maximum implies that the term in braces is zero; hence the right-hand side of (A6) is also zero and (A3) has been shown to hold.

## REFERENCES

K. J. Arrow, *Essays in the Theory of Risk-Bearing*, Chicago 1971.

J. R. Hicks, *Critical Essays in Monetary Theory*, Oxford 1967.

———, "Liquidity," *Econ. J.*, Dec. 1962, *72*, 787–802.

J. M. Keynes, *Treatise on Money*, Vol. II, London 1950.

M. Rothschild and J. E. Stiglitz, "Increasing Risk I: A Definition," *J. Econ. Theory*, Sept. 1970, *2*, 225–43.

——— and ———, "Increasing Risk II: Its Economic Consequences," *J. Econ. Theory*, Mar. 1971, *3*, 66–84.

S. C. Tsiang, "The Precautionary Demand for Money: An Inventory Theoretical Analysis," *J. Polit. Econ.*, Jan.-Feb. 1969, *77*, 99–117.

# Inflationary Expectations and Momentary Equilibrium

By LEWIS JOHNSON*

By introducing the Fisherian distinction between real and nominal rates of return into a conventional static macro-equilibrium model, Robert Mundell illustrated how macro theory could be modified to analyze the impact of inflationary expectations on equilibrium. Although details of the analysis permit considerable elaboration of the model studied by Mundell, the basic analytical framework has become standard. This framework maintains the usual specification of asset market equilibrium: that in momentary equilibrium all agents must be satisfied with the current composition of their portfolios.[1] This may be called a *portfolio balance* (*PB*) specification of asset market equilibrium.

It is with this adaptation of the *PB* model to an analysis of inflationary expectations that I wish to take issue. Unless transacting in asset markets is costless, wealth holders will not instantaneously adjust imbalances in the composition of their portfolios. But the introduction of inflationary expectations means that they must anticipate that the real value of their cash balances will be declining over time due to inflation. Wealth holders will plan to readjust their portfolios over time, taking into account this anticipated outflow of real balances. In Section I, a stock adjustment model of portfolio behavior is proposed. Asset markets are defined to be in equilibrium when the adjustment plans of individual wealth holders are consistent for the current moment. This will be called a portfolio adjustment (*PA*) model of asset market equilibrium. The *PB* model emerges as a special case in which portfolios are instantaneously adjusted. In Section II, the comparative statics of an increase in the expected rate of inflation are analyzed for full employment and rigid wages versions of both *PA* and *PB* models. It is shown that these alternative specifications of asset market equilibrium can lead to qualitatively different comparative statics results.

It is legitimate to ask why we should concern ourselves with this issue between two static theories. After all, static analysis is inherently incomplete as it is not possible to account for the impact of steadily accumulating capital (arising from nonzero investment) in a static framework. To discuss "long-run" equilibrium properties of an economy in a static framework is highly tenuous, at best. Nonetheless, the static framework is invaluable in defining positions of momentary equilibrium which partly determine the system's position at each moment in time. The comparative statics analysis of the momentary equilibrium will determine in part how variables will move as the parameters of the momentary model change in accord with the equations of motion of the dynamic system. In particular, in a model in which inflationary expectations are learned adaptively, the amount by which prices must change (in momentary equilibrium) in response to an autonomous rise in the expected rate of inflation will be critical for the stability of the dynamic system. The implications of the *PB* and *PA* models for this problem of dynamic stability will be discussed in Section III.

## I. A Portfolio Adjustment Model of Momentary Equilibrium

Implicit in the analysis of a monetary economy is the acknowledgement that frictions and uncertainties exist, for these provide the rationale for holding money balances. The existence of costs of transacting

* Economist, special studies section, Division of Research and Statistics, Federal Reserve Board. I am indebted to my colleagues Stephen LeRoy, Steven Salop, Roger Waud, and to an anonymous referee for helpful comments and suggestions, but I am solely responsible for any remaining errors. The views expressed herein are solely my own and do not necessarily represent the views of the Board of Governors of the Federal Reserve System.

[1] The term momentary equilibrium will be used to mean that the markets under discussion are (momentarily) cleared.

in capital and bond markets implies that wealth holders will formulate plans determining an optimal (time) rate of acquisition of money balances, titles to capital, and bonds in the face of a divergence between desired and actual holdings. Only in the extreme case of zero transactions costs (implying that it is *feasible* to make instantaneous transactions) will wealth holders always instantly adjust actual to desired holdings.

Clearly, the plans of many independent agents are not in general going to be consistent. It is the role of markets to adjust prices until the plans for current action are made consistent. Momentary equilibrium is a state in which the plans for the current moment are consistent over all agents. This does not require that each agent intends to hold his portfolio in its current composition. It merely requires that those who enter the market to dispose of stocks at some (time) rate find buyers wishing to acquire stocks at that same rate. Aggregate flow excess demand must be zero in momentary equilibrium. For a full employment model, then, we may characterize equilibrium by the condition that the flow excess demand for goods and the flow excess demand for money are both zero.

Consider first the flow excess demand for money. As the simplest case, assume that desired holdings of real balances $m$ are a function of real income $y$ and the nominal rate of interest $r$. Assume also that the chosen plan for accumulation is to acquire real balances at a rate proportional to the discrepancy between desired and actual holdings. Then the desired flow of real balances $Dm^d$ may be written:[2]

$$(1) \qquad Dm^d = \lambda(L(y, r) - m)$$

It is not necessary, however, for the individual wealth holder to enter the market to the extent indicated by (1) to achieve the desired adjustment. He may anticipate an inflow of real balances in the form of a) a share of government transfers and b) appreciation of his current money balances due to deflation. This anticipated inflow of real balances $Dm^e$ will partially accommodate his desired adjustment. Consequently, the market flow excess demand for real balances is:

$$(2) \quad Dm^d - Dm^e = \lambda(L(y, r) - m) - Dm^e$$

By definition:

$$Dm^e = (\theta - q)m$$

where $$\theta = \left(\frac{DM}{M}\right)^e, \qquad q = \left(\frac{Dp}{p}\right)^e$$

Consequently, momentary equilibrium implies:

$$(3) \quad \lambda(L(y, r)) - (\lambda + \theta - q)m = 0$$

This constitutes a portfolio adjustment specification of asset market equilibrium, at least in this case with the goods market assumed to be simultaneously cleared. This is quite distinct from the conventional liquidity preference ($PB$) model which requires that the stock of real balances equal the (stock) demand for real balances:

$$(3') \qquad L(y, r) - m = 0$$

For the analysis of most comparative statics problems it is assumed that $\theta$ and $q$ are both zero (or at least equal). In this case it is evident that the $PA$ and $PB$ models are the same. Since wealth holders (in this case) expect no flow supply of real balances, equilibrium requires that flow demand be zero, which will only occur if actual stock holdings equal desired stock holdings. Obviously, this requirement must be satisfied in long-run equilibrium (steady state). Thus the distinction can only arise in the analysis of momentary equilibrium. Also, it is clear that in the case that $\lambda$ tends to infinity there must be stock equilibrium (3′) to satisfy flow equilibrium (3). Again the models are identical.

However, if we wish to admit the existence of transactions costs in securities markets ($\lambda$ finite) and at the same time attempt to analyze the impact of an *autonomous* change in inflationary expectations (as in the articles by William Gibson, Robert Mundell, Edmund Phelps, Frank Steindl, and Roger Waud), it is evident that the standard $PB$ model is inadequate. Since the expectation

[2] The symbol $D$ will be used to denote the operator $d/dt$.

of inflation implies an anticipated depreciation of the real value of money holdings (a negative flow "supply"), the *PA* model must be used to describe equilibrium.

Goods market equilibrium is characterized by the condition that planned investment equal planned savings:

$$I(\rho) - S(y, \rho, m) = 0 \tag{4}$$

with $I'(\rho) < 0$, $S_1 > 0$, $S_2 \geq 0$, $S_3 \leq 0$; where $I(\rho)$ is planned investments, $S(y, \rho, m)$ is desired savings, and $\rho$ is the real rate of interest: $\rho = r - q$. Equations (3′) and (4) are the standard *PB* model of momentary equilibrium, and equations (3) and (4) constitute the proposed *PA* model. The impact of inflationary expectations will be analyzed for each of these models for two extreme cases: a) full employment ($y$ fixed) and b) variable real income (money wage fixed).

## II. The Impact of Inflationary Expectations on Momentary Equilibrium

### A. *Full Employment*

Consider first the effects of an autonomous increase in the expected rate of inflation on the full employment *PB* model. The simplest version of this analysis is as follows. At the initial nominal rate, the anticipation of inflation reduces the real cost of investing and simultaneously lowers the inducement to save. The resulting excess demand for goods forces prices to rise, lowering the existing stock of real balances. Wealth holders attempt to sell bonds in order to acquire cash balances which forces up the nominal rate of interest. In the absence of wealth effects ($S_3 = 0$), only the initial real rate is consistent with equilibrium in goods markets. Consequently, the nominal rate must rise by the amount of anticipated inflation and prices must rise sufficiently to reduce real balances to the level demanded at this higher nominal interest rate.

If there is a significant wealth effect ($S_3 < 0$) it is only necessary to note that the reduction of real balances (due to higher prices) has an effect on the goods market, insofar as it stimulates savings (the real balances effect). This requires a somewhat lower real rate of interest to clear the goods market. Thus the real rate falls, and the nominal rate rises by less than the expected rate of inflation. These are the results derived Mundell.

Analytical expressions for the impact of inflationary expectations on the interest rates and on prices may be obtained by differentiating (3′) and (4) with respect to $q$ and solving for $dr/dq$, $d\rho/dq$, and $(1/p) \cdot (dp/dq)$.[3] The solutions are:

$$\frac{dr}{dq} = \frac{-(I' - S_2)}{L_2 S_3 - (I' - S_2)} > 0 \tag{5}$$

$$\frac{d\rho}{dq} = \frac{dr}{dq} - 1 = \frac{-L_2 S_3}{L_2 S_3 - (I' - S_2)} < 0 \tag{6}$$

$$\frac{1}{p}\left(\frac{dp}{dq}\right) = \frac{L_2(I' - S_2)}{m[L_2 S_3 - (I' - S_2)]} > 0 \tag{7}$$

Consider the impact on prices. Prices must rise in order to reduce the level of real balances to that which will be desired at the new higher nominal rate of interest. But suppose that portfolios are not instantly readjusted. Then the expectation of inflation implies that people expect the real value of their cash balances to dwindle over time. If the period of adjustment is of precisely the right length, individual wealth holders need not enter the market at all to readjust their portfolios because their holdings of real balances will be reduced automatically by inflation. Consequently, no "once over" rise in prices will be required to maintain equilibrium. This anticipation effect is accounted for by the *PA* model. For this model, differentiate (3) and (4) with respect to $q$, and solve to get:

$$\frac{dr}{dq} = \frac{-[S_3 m + \lambda(I' - S_2)]}{\lambda[L_2 S_3 - (I' - S_2)]} > 0 \tag{8}$$

$$\frac{d\rho}{dq} = \frac{dr}{dq} - 1 = \frac{-S_3(m + \lambda L_2)}{\lambda[L_2 S_3 - (I' - S_2)]} \gtrless 0 \tag{9}$$

$$\frac{1}{p}\left(\frac{dp}{dq}\right) = \frac{(I' - S_2)(m + \lambda L_2)}{m\lambda[L_2 S_3 - (I' - S_2)]} \gtrless 0 \tag{10}$$

[3] For convenience of exposition, it will be assumed that the initial rate of monetary expansion and the initial expected rate of inflation are both zero.

According to (8) the nominal rate of interest must rise, as in the *PB* model. As may be seen from (9) and (10), however, although prices must again move in the opposite direction from the real rate (so long as $S_3 \neq 0$), they may either rise, fall, or remain unchanged. Thus the anticipation effect may be sufficiently strong to reverse the qualitative results predicted by the standard (*PB*) model.[4] The initial real interest rate and prices will be just maintained in the new equilibrium if

$$m + \lambda L_2 = 0$$

or

$$1/\lambda = -\frac{L_2}{m}$$

For convenience, define the interest elasticity of demand for money $\epsilon$ to be $-L_2/m$, despite other usage. Thus, if the average adjustment period $1/\lambda$ is equal to the elasticity of demand for money, the expected inflation will lead wealth holders to anticipate that their cash balances will be automatically adjusted at the desired rate. The elasticity of demand may be interpreted as the length of time expected for inflation to reduce the real value of cash balances to the level that would be desired if the nominal rate rises by exactly the rate of inflation. If this equals the desired lag in adjustment $1/\lambda$, prices and real rates will be unchanged. If people plan to make the adjustment in less time than $\epsilon$, then the standard *PB* model's results will be qualitatively correct: prices must rise and the real rate will fall. If they plan their adjustments to require more time than $\epsilon$, then prices will fall, and the real rate will rise. Thus the two models may have qualitatively different predictions.

## B. *The Variable Income Model*

In response to the demonstration by Phelps that anticipated inflation necessarily reduces "feasible" welfare, Waud argued that if the system were initially in a position of less than full employment (due to rigid money wages), anticipated inflation would provide a stimulus to aggregate demand, resulting in a higher equilibrium level of income; consequently anticipated inflation might well increase welfare. This increase in output is possible because a rise in prices reduces the real wage given the money wage. The analysis is much the same as that for the full employment case except that any movement in prices is accompanied by a change in income in the same direction. Thus for the *PB* model, an increase in anticipated inflation implies a rise in the nominal rate of interest, prices, and income and a fall in the real rate of interest (even in the case $S_3=0$, since income has risen).

In the *PA* version of the variable employment model, these results may not hold. The reasoning is the same as for the full employment case. Individual agents perceive that the inflation is reducing their real balances over time, and if the desired reduction is achieved rapidly enough, the system will remain in equilibrium at the initial level of prices. In fact if the reduction is proceeding too rapidly, prices may actually fall. Income is positively related to the price level because of the fixed-money wage, so income will rise (as in Waud's analysis) under the same conditions that prices will rise; specifically, income will rise due to inflationary expectations if and only if the average period of portfolio adjustment $1/\lambda$ is less than the interest elasticity of demand for money $\epsilon$. The reader may readily confirm these results algebraically.

## III. Stability

If the model implies a "once over" rise in prices in response to inflationary expectations, the question of how much they rise may be crucial for the stability of the system. Although it is tenuous to discuss stability without specifying a complete dynamic model, we can perhaps arrive at a few tentative conclusions. Suppose that expectations are revised adaptively such that the current error between expected and actual rates of inflation is corrected at a rate proportional (by a factor of $\beta$) to that error, i.e.:

[4] It should be noted that this perverse case strains the interpretation of the finite rate of adjustment arising from the existence of transactions costs in asset markets. A more thorough analysis than is currently available of portfolio behavior with transactions costs and inflationary expectations would be desirable.

$$(11) \qquad Dq = \beta(Dp/p - q)$$

It is readily possible that this adaptive expectations process will permit a self-sustaining expectations inflation, since an increase in the expected rate of inflation may itself cause prices to rise. To avoid instability, the rise in prices must not be sufficiently great to validate the increased inflationary expectations. Specifically, equation (11) will be a stable process if and only if:

$$(12) \qquad \frac{\partial(Dp/p)}{\partial q} < 1$$

The left-hand side of this inequality may be obtained by differentiating $p = p(q, M)$ with respect to time, substituting from (11) for $Dq$ and solving for the rate of inflation. Taking the partial with respect to $q$ of the resulting expression for $Dp/p$, inequality (12) may be arranged to yield:

$$(13) \qquad \frac{1}{p}\frac{dp}{dq} < \frac{1}{\beta}$$

This is the necessary and sufficient condition for stability of the expectations process. The gradual readjustment of actual portfolio composition to desired composition in the *PA* model will reduce the amount by which prices must rise in response to an increase in anticipated inflation, thus tending to stabilize the system. To see this, examine the "stability condition" (13) for both models.

Substituting from equation (7) into (13), the condition for stability of the PB model becomes:

$$(\beta L_2 + m)(I' - S_2) - mL_2S_3 < 0$$

As the wealth effect on savings is usually thought to be small, an interesting sufficient condition for this inequality to be satisfied is obtained by requiring that the first term be negative, which implies:[5]

$$(14) \qquad -\frac{L_2}{m} \equiv \epsilon < \frac{1}{\beta}$$

Similarly, (by (10) and (13)) the condition for stability of the PA model may be shown to be:

$$\beta[(m + \lambda L_2) + m\lambda][I' - S_2] - m\lambda S_3L_2 < 0$$

and the sufficient condition corresponding to (13) is:

$$(15) \qquad \epsilon < \left(\frac{1}{\beta}\right) + \left(\frac{1}{\lambda}\right)$$

Gradual portfolio adjustment clearly tends to stabilize the system, entering in precisely the same way as the average lag in expectations formation. For any values of $\epsilon$ (finite) and $\beta$, the system will be stable if portfolio adjustment is sufficiently slow. This is clear from equation (10) alone, since $\lambda$ may be small enough for prices to fall (in momentary equilibrium) in response to an increase in inflationary expectations.

A related point to the difference of the stability requirements of the two models is the efficacy of stabilization policy via the rate of monetary expansion $\theta$. It is evident by inspection of (3′) and (4) that tying $\theta$ inversely to the anticipated rate of inflation will have no impact on stability of the system in the *PB* model.[6] This is not true of the *PA* model however. In particular, the immediate impact of an increase in $\theta$ on the equilibrium price level is:

$$\frac{1}{p}\frac{dp}{d\theta} = \frac{(I' - S_2)}{\lambda[L_2S_3 - (I' - S_2)]} > 0$$

If $S_3$ is small, this is approximately $1/\lambda$. Thus it would be possible to design a stabilization policy by tying $\theta$ to $q$ if the length of time economic agents choose to readjust their portfolios is appreciable. The derivation of such a policy for a completely specified dynamic model is shown in my forthcoming paper.

[5] This is precisely the stability criterion first derived by Phillip Cagan for the single equation model discussed here and later proved to be necessary for stability of an equilibrium monetary growth model by Miguel Sidrauski. To see how this condition is modified in a model with gradual price adjustment, see the paper by Michael Hadjimichalakis (1971).

[6] As shown by Hadjimichalakis (1973), an alternative specification of monetary policy may be effective. Also, Robert MacKay has demonstrated that policy may be effective if asset market equilibrium accounts for the flow of nominal money, a step in the direction of the *PA* model.

## IV. Summary and Conclusions

The conventional *IS-LM* apparatus with a portfolio balance model of asset market equilibrium is adequate for the analysis of most problems in momentary equilibrium. It is, however, ill-suited for the analysis of the impact of an autonomous change in expectations on momentary equilibrium, except in the extreme case in which portfolios are instantaneously (costlessly) readjusted. The standard model ignores the anticipation effect: that wealth holders perceive that the real value of their cash balances will be declining over time, thus assisting them in readjusting their portfolios toward smaller money holdings. The portfolio adjustment theory of asset market equilibrium does take account of this effect and results in qualitatively different conclusions in the analysis of momentary equilibrium.

In addition to pointing out the different comparative statics results of a model incorporating gradual portfolio adjustment, the model was extended to examine stability and stabilization policies for an adaptive expectations process. While this framework does serve its purpose to illustrate clearly the nature and significance of the issues involved, it is evident that the assumed dynamics are somewhat implausible. In particular, it would seem to be farfetched to assume that portfolio adjustment requires an appreciable interval of time while goods markets are assumed to adjust instantaneously. However, the important result that gradual portfolio adjustment tends to stabilize the system (to the same degree as does gradual expectations formation) is retained when the portfolio adjustment process is imbedded in a general monetary growth framework, as is shown in my forthcoming paper.

## REFERENCES

P. Cagan, "The Monetary Dynamics of Hyperinflation," in M. Friedman, ed., *Studies in the Quantity Theory of Money,* Chicago 1956, 25–117.

W. E. Gibson, "Interest Rates and Monetary Policy," *J. Polit. Econ.,* May 1970, *78,* 431–55.

M. G. Hadjimichalakis, "Equilibrium and Disequilibrium Growth with Money—The Tobin Models," *Rev. Econ. Stud.,* Oct. 1971, *38,* 457–79.

———, "On the Effectiveness of Monetary Policy as a Stabilization Device," *Rev. Econ. Stud.,* Oct. 1973, *40,* 561–70.

L. Johnson, "Portfolio Adjustment and Monetary Growth," *Rev. Econ. Stud.,* forthcoming.

R. J. Mackay, "Stein and Nagatani on Stabilization Policies in a Growing Economy," *Rev. Econ. Stud.,* Oct. 1973, *40,* 571–78.

R. A. Mundell, "Inflation and Real Interest," *J. Polit. Econ.,* June 1963, *71,* 280–83.

E. S. Phelps, "Anticipated Inflation and Economic Welfare," *J. Polit. Econ.,* Feb. 1965, *73,* 1–17.

M. Sidrauski, "Inflation and Economic Growth," *J. Polit. Econ.,* Dec. 1967, *75,* 796–810.

F. G. Steindl, "Price Expectations and Interest Rates," *J. Money, Credit, Banking,* Nov. 1973, *5,* 939–49.

R. N. Waud, "Inflation, Unemployment, and Economic Welfare," *Amer. Econ. Rev.,* Sept. 1970, *60,* 631–41.

# The Borrower-Lender Relationship, Competitive Equilibrium, and the Theory of Hedonic Prices

*By* ERNST BALTENSPERGER*

In a series of recent articles, Vernon Smith (1972b; see also 1971a, b, 1972a) has presented a very interesting analysis of the lender-borrower relationship, which emphasizes that the debtor's equity capital plays the role of an external economy to the lender, with the result that the traditional competitive equilibrium solution is not Pareto optimal. Smith considers a situation where a lender with initial wealth $w_0$ has the option of either buying government bonds with a certain return $r$, or giving a risky loan $z$ at a contractual rate of interest $r^*$ to an individual borrower or firm. The borrower uses his loan to finance a risky investment project yielding an uncertain return $\theta$ (assumed independent of the amount invested). The borrower also has an initial wealth $W_0$, which he invests partly in his risky investment project (his "firm") and partly in the form of riskless government securities at rate $r$. Since the borrower's risky investment activity (his firm) is assumed to be incorporated with limited liability, so that his private holdings of riskless bonds are protected from lender claims, the lender's expected utility from extending a loan clearly becomes dependent on the debtors' equity decision, i.e., on the way in which the debtor divides his personal wealth $W_0$ into investment in the risky project $y$ (his equity in his firm) and investment in riskless bonds (his personal, nonliable assets). The higher $y$, the lesser is (given the size of the loan $z$) the risk that the debtor will default on his loan, and thus the higher is the lender's expected utility. The debtor's equity capital acts, as Smith expresses it, as an external economy to the lender. The existence of this external effect implies that the standard competitive equilibrium solution, as summarized by his (1972b) equations (12) and (13), is not Pareto efficient. Therefore, he suggests that free competitive negotiations between borrowers and lenders may produce arrangements which look quite different from what is normally expected to characterize a "competitive equilibrium." More specifically, he links his analysis to the discussion about credit rationing (see Dwight Jaffee-Franco Modigliani, 1969, and Jaffee, 1971), and points out that a quota or rationing system can generate a solution which is consistent with the conditions for Pareto efficiency, if the quotas are set such that their shadow values are equal to the appropriate tax or subsidy. But he does not have a theory which explains how these quotas are determined as the outcome of a market process.

Fundamental to Smith's whole analysis is the behavioral assumption that both lender and borrower optimize based on the premise that the borrower's interest payment to the lender is independent of the debtor's equity capital $y$, and thus of his risk of default. The borrower takes the contractual interest rate $r^*$ as a datum and optimizes based on the assumption that he can obtain any amount of credit at this rate. The borrower has no inducement to increase his equity capital $y$ towards the Pareto optimal level, since his interest payment $r^*z$ depends, for any parametrically given value of the price $r^*$, on $z$ only, independent of $y$.

This paper discusses a model of the competitive process based on an alternative assumption about competitive behavior, one which is in my opinion more justified, and which leads to a Pareto-efficient solution. In contrast to Smith's rationing scheme, this solution is perfectly consistent with "what is usually expected to characterize a competitive equilibrium." In the context of a credit market system as discussed here, the assumption that the transactors optimize subject to a parametrically given contractual

* Associate professor of economics, Ohio State University.

interest rate $r^*$ does not seem justified, even in an atomistic environment. The assumption of price taking is, of course, a very standard behavioral assumption in economic theory, usually used as equivalent with "competitive behavior," and normally justified by assuming that the seller or buyer in question is but one among a large number of transactors, so that his individual behavior does not appreciably affect the market as a whole nor the market price. Clearly, this justification for price taking, if applied to the contractual interest rate $r^*$ as in Smith's model, is invalid. Every borrower, however small and insignificant he is relative to the market as a whole, knows that, given his equity capital $y$, he cannot borrow unlimited amounts at the same (contractual) interest rate, since an increase in his loan (or, alternatively, a change in his equity capital), *ceteris paribus*, will change his risk of default and therefore the "quality" of the loan. Recognition of this has nothing to do with presence or absence of competition. Even in an atomistic environment, a borrower is not reasonable if he does not take into account that the contractual rate $r^*$ must contain a risk premium which reflects the characteristics and riskiness of the individual borrower, and thus his equity decision, rather than general demand and supply conditions for the market as a whole. A credit transaction has two dimensions or "characteristics": dollar amount (or "quantity") and riskiness (or "quality," depending on $y$, relative to $z$). Both parties are perfectly well aware of this (and they know that the other side is aware of it), and thus should not be expected to optimize on the premise that the other side is *not* aware of it. Consequently, I suggest that the loan should be treated as a "good" with two characteristics. There is a perfect formal analogy between the problem under discussion here and the model of hedonic prices employed by Sherwin Rosen to discuss product differentiation in a competitive environment (which is based on the well-known work on consumer theory by Kelvin Lancaster, Gary Becker, and Hendrik Houthakker, where goods are treated as collections of utility generating characteristics).

Thus, I object to Smith's description of the conventional competitive model, since it is equivalent to assuming that the market participants optimize based on the assumption that price is independent not only of quantity, but also of quality. There is nothing in the traditional competitive model requiring this. It is shown that if this assumption is abandoned, and the loan is treated as a good with two characteristics, a competitive (atomistic) environment can be expected to produce a Pareto-efficient solution quite consistent with what is usually expected to characterize a competitive equilibrium (in other words: a situation where the appropriate "charges" or "subsidies" are imposed quite naturally as the outcome of a market process).

In Section I below, Smith's model will be briefly summarized. In Section II, a competitive model which does take account of the preceding discussion will be presented. Section III contains a brief summary.

## I

In the Smith model, the borrower's terminal wealth $W$ is

$$(1+r)(W_0-y) + (1+\theta)(y+z) - (1+r^*)z$$

in the event of nondefault. This occurs when $(1+\theta)(y+z)$, the end of period value of his investment project (or firm), exceeds his debt to the lender $(1+r^*)z$, or in other words,

$$\theta > [(1+r^*)z/(y+z)] - 1 \equiv \theta^*$$

On the other hand, in the event of default, the end of period value of his personal holdings of nonliable bonds equals $(1+r)(W_0-y)$, i.e., $\theta<\theta^*$.

The borrower is assumed to maximize his expected utility

$$(1)\quad B = \int_{\theta^*}^{\infty} U[(1\,r)(W_0-y) + (1+\theta)(y+z) - (1+r^*)z]f(\theta)d\theta + \int_{-1}^{\theta^*} U[(1+r)(W_0-y)]f(\theta)d\theta$$

where $U(W)$ denotes his (strictly concave) utility function for terminal wealth, and $f(\theta)$ his subjective probability density over the rate of return $\theta$. Maximizing $B$ with respect to $y$ and $z$, for given $r^*$, $r$, and $W_0$, determines the borrower's optimal equity capital $y$ and his loan demand $z$, as a function of $r^*$, $r$, and $W_0$.

The lender's terminal wealth $w$, on the other hand, is equal to

$$(1 + r)(w_0 - z) + (1 + r^*)z$$

in the event of nondefault ($\theta > \theta^*$), but only

$$(1 + r)(w_0 - z) + (1 + \theta)(y + z)$$

in the case of default ($\theta < \theta^*$).[1] The lender is assumed to maximize his expected utility

$$(2) \qquad L = \int_{\theta^*}^{\infty} V[(1 + r)(w_0 - z) + (1 + r^*)z]g(\theta)d\theta + \int_{-1}^{\theta^*} V[(1 + r)(w_0 - z) + (1 + \theta)(y + z)]g(\theta)d\theta$$

where $V(w)$ denotes his utility function for terminal wealth, and $g(\theta)$ his subjective density function over $\theta$. Maximization of $L$ with respect to $z$, subject to given values of $r^*$, $r$, $w_0$, and $y$, yields the lender's optimal loan supply $z$.

Clearly, a parametric change in $y$ changes the value of the lender's expected utility $L$:

$$(3) \qquad \frac{\partial L}{\partial y} = \int_{-1}^{\theta^*} V'(\cdot)(1 + \theta)g(\theta)d\theta > 0$$

and thus leads, *ceteris paribus*, to a change in the lender's loan supply. "The borrower's equity in an investment is an external economy to the lender," as Smith expresses it (his Proposition 1, p. 479, 1927b). The existence of this external effect leads to the conclusion that the traditional competitive market equilibrium is not Pareto efficient, as noted in the introduction.

## II

Smith's result is obtained because of the special (and, as argued here, objectionable) way in which the behavior of the market participants is specified. More specifically, because it is assumed that both borrower and lender optimize based on the premise that the contractual interest rate $r^*$ and thus the borrower's interest payment to the lender, is independent of his equity capital $y$, and thus of his risk of default. This is equivalent to assuming that price is independent of quality, and does not seem justified, even in an atomistic environment. I may expect that I can buy "any" amount of Coca Cola in the supermarket (i.e., any amount within a very large range around my current purchase, say, 100 bottles instead of my normal purchase of 1) without exerting any noticable effect on the market price, because I am but one of millions of purchasers of Coca Cola. However, even if I am but one among millions of (similar) credit takers, I will never expect that I can get "any" amount of credit at an unchanged contractual rate of interest, unless I keep default risk (the "quality" of the loan) constant by appropriately adjusting equity capital $y$ (or in some other way).[2] To assume competitive behavior in the sense of having many relatively small market participants whose behavior taken individually does not appreciably affect total market excess demand ("general market conditions") should not be equated with price taking in the sense of accepting $r^*$ as a datum (independent of $y$).

This paper suggests treating the loan as a good with two characteristics, $z$ and $y$ (size and quality). Consequently, we specify a price function $R = R(z, y)$, which relates the borrower's payment to the lender $R$ (defined as including the repayment of capital) to the

[1] To simplify the presentation, it is assumed that the lowest possible value of $\theta$ is not below $-1$, i.e., the lowest possible terminal value of the lender's investment project not less than zero. If lower values for $\theta$ were admitted, we would have to include a third term in (2), since in this case the lender's terminal wealth would be given by $(1+r)(w_0-z)$ alone (because he is not liable for the borrower's debt to somebody else).

[2] For example, via putting up otherwise nonliable wealth as collateral. This would be an obvious extension of the model.

characteristics $z$ and $y$.[3] At any point in time, the market reveals an $R(z, y)$ function, which can be inferred from observed prices for different $z/y$ packages. Presumably, $R_z>0$ and $R_y<0$: an increase in $z$ implies a larger total payment $R$, while an increase in $y$, given $z$, represents an increase in the quality of the loan, and thus implies a lower compensation $R$. The transactors treat this price function as parametric to their decisions. This is the meaning of competition in this context. In other words, each transactor is small enough relative to the market as a whole that his individual actions have no recognizable effect on the $R(z, y)$ function exhibited by the market. An equilibrium price function is determined by market forces in the same way as equilibrium price is determined in the usual competitive model: an equilibrium value for $R(z, y)$ is reached if buyers and sellers are perfectly matched at each $z/y$ combination. For a discussion of the determination of such an equilibrium price function, see Rosen and Louis Court.

The borrower's and lender's expected utility functions now are

$$(4)\quad B = \int_{\theta^*}^{\infty} U[(1+r)(W_0 - y) + (1+\theta)(y+z) - R(z,y)]f(\theta)d\theta + \int_{-1}^{\theta^*} U[(1+r)(W_0 - y)]f(\theta)d\theta$$

and

$$(5)\quad L = \int_{\theta^*}^{\infty} V[(1+r)(w_0 - z) + R(z, y)]g(\theta)d\theta + \int_{-1}^{\theta^*} [V(1+r)(w_0 - z) + (1+\theta)(y+z)]g(\theta)d\theta$$

where $\theta^* = [R(z, y)/(y+z)] - 1$ again is the value of $\theta$ below which the borrower defaults on his loan.

Treating $R(z, y)$ parametrically, the borrower determines his optimal $z$ and $y$ (size and quality of loan). His optimality conditions are

$$(6)\quad \frac{\partial B}{\partial z} = \int_{\theta^*}^{\infty} U'(\cdot)[1+\theta - R_z]f(\theta)d\theta = 0$$

$$(7)\quad \frac{\partial B}{\partial y} = \int_{\theta^*}^{\infty} U'(\cdot)[\theta - r - R_y]f(\theta)d\theta - \int_{-1}^{\theta^*} U'(\cdot)(1+r)f(\theta)d\theta = 0$$

Similarly, the lender determines, conditional upon $R(z, y)$, his desired values of $z$ and $y$, i.e., the size and quality of the loan he would like to make, given the market's $R(z, y)$:[4]

$$(8)\quad \frac{\partial L}{\partial z} = \int_{\theta^*}^{\infty} V'(\cdot)[R_z - 1 - r]g(\theta)d\theta + \int_{-1}^{\theta^*} V'(\cdot)[\theta - r]g(\theta)d\theta = 0$$

$$(9)\quad \frac{\partial L}{\partial y} = \int_{\theta^*}^{\infty} V'(\cdot)R_y g(\theta)d\theta + \int_{-1}^{\theta^*} V'(\cdot)(1+\theta)g(\theta)d\theta = 0$$

A competitive market equilibrium is reached if the $R(z, y)$ function is such that there results complete consistency between borrowers and lenders. See Rosen for a discussion of the price function as the joint envelope of a family of indifference surfaces for borrowers and lenders. Of course, the existence of a whole equilibrium price *function*, i.e., the coexistence of different $z/y$ combinations in equilibrium, requires the existence of heterogeneity among borrowers and/or lenders (for example, a distribution of $U$-functions, and/or distribution of $V$-functions). If all borrowers as well as all lenders were identical, all borrowers and all

[3] Note that $z$ and $y$ together determine $\theta^*$, the critical value of $\theta$ below which the debtor defaults on his loan. The "quality" of the loan, of course, also depends on the density functions over $\theta$, which we treat as given here. In principle, we can think of the parameters of these functions as additional characteristics of the loan.

[4] In addition, the appropriate second-order conditions for a maximum must be satisfied.

lenders would always end up with identical decisions, so that there would be only one equilibrium $z/y$ combination. Consequently, the market then could not provide the information necessary to infer a complete $R(z, y)$ function. A discussion of how an equilibrium price function is reached, starting from a disequilibrium situation, is clearly beyond the scope of this paper. However, it can easily be seen that such an equilibrium, if obtained, satisfies all the conditions for Pareto optimality. Using the shorthand notation employed by Smith (1972b, p. 481), we can rewrite the borrower's and lender's objective functions (4) and (5) as $B=B(u, z, y/R(z, y), r)$ and $L=L(v, z, y/R(z, y), r)$, with $u=W_0-y$ and $v=w_0-z$. Independent maximization by both sides yields the following marginal conditions (equivalent to (6)–(9))

$$B_u = B_y, \qquad B_z = 0 \tag{10}$$

and

$$L_v = L_z, \qquad L_y = 0 \tag{11}$$

which is perfectly consistent with the condition for Pareto efficiency (as summarized by Smith's (14), 1972b, p. 481).

## III

I have discussed a competitive model of the credit market which does explicitly take into account that the credit transaction has two characteristics: size and riskiness, or "quality." It has been argued that it is reasonable to assume that market participants do explicitly take this into account in their optimization procedures; and it has been shown that a model of the competitive process where this is the case does lead to a Pareto-efficient solution which is perfectly consistent with what is usually expected to characterize a competitive equilibrium (in contrast to the competitive model discussed by Smith).

## REFERENCES

G. S. **Becker**, "A Theory of the Allocation of Time," *Econ. J.*, Sept. 1965, *75*, 493–517.

L. M. **Court**, "Entrepreneurial and Consumer Demand Theories for Commodity Spectra," *Econometrica*, Apr. 1941, *9*, 135–62, Part I; July–Oct. 1941, *9*, 241–97, Part II.

H. S. **Houthakker**, "Compensated Changes in Quantities and Qualities Consumed," *Rev. Econ. Stud.*, 1952, *19*, 155–64.

D. M. **Jaffee**, *Credit Rationing and the Commercial Loan Market*, New York 1971.

——— and F. **Modigliani**, "A Theory and Test of Credit Rationing," *Amer. Econ. Rev.*, Dec. 1969, *59*, 850–72.

K. J. **Lancaster**, "A New Approach to Consumer Theory," *J. Polit. Econ.*, Apr. 1966, *74*, 132–56.

S. **Rosen**, "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *J. Polit. Econ.*, Jan./Feb. 1974, *82*, 34–55.

V. L. **Smith**, (1971a) "The Borrower-Lender Contract under Uncertainty," *Western Econ. J.*, Mar. 1971, *9*, 1075–78.

———, (1971b) "Investment Decision, Uncertainty, and the Incorporated Entrepreneur," in J. Quirk and A. Zarley, eds., *Papers in Quantitative Economics*, II, Kansas City 1971.

———, (1972a) "Default Risk, Scale, and the Homemade Leverage Theorem," *Amer. Econ. Rev.*, Mar. 1972, *62*, 66–76.

———, (1972b) "A Theory and Test of Credit Rationing: Some Generalizations," *Amer. Econ. Rev.*, June 1972, *62*, 477–83.

# The Borrower-Lender Relationship

*By* VERNON L. SMITH*

One frequently encounters the belief that there are pervasive "imperfections in the capital market" (see George Stigler), and the widespread practice of lender rationing in loan markets is often implicated as providing the evidence in support of this proposition (see Dwight Jaffee and Franco Modigliani). I have long thought this to be a simplistic proposition, and my papers on the borrower-lender contract are an attempt to interpret the phenomenon of lender or capital rationing as an instance in which competitive markets were doing their job but where the "commodity" had unusual or nonstandard properties. The loan contract has the nonstandard characteristic that the lender's gain from the exchange depends upon a host of borrower controlled activities. Thus the lender wants to know how the proceeds of the loan are to be invested, the level of borrower equity, and the amounts and terms of any other loans contracted.

A formal way to treat such a commodity is to enrich its price expression to correspond to its enriched space of characteristics. There are several well-known examples of this formalism. By treating externalities as commodities and giving them prices, one recovers the competitive efficiency properties otherwise lost. Efficiency in the presence of uncertainty is no problem, formally, if one simply invents state-contingent claims. In this spirit, a natural way to treat the loan contract is to let its market price (the nominal interest rate) be a function of the borrower's equity, or any other relevant characteristics, as well as the amount of the loan. This approach is so natural that I did it myself in several pages of the working notes for the papers cited by Ernst Baltensperger. The published version did not develop this line of analysis because I thought it finessed the issue of capital rationing in the borrower-lender contract which was my principal interest. Also we know that, in general, competitive loan markets could improve on the Baltensperger price function if borrowers would issue state-contingent claims. The proper full formal treatment of the loan contract is to see that it has as many dimensions as contingent states, $\theta$, and that claims contingent on each such state must have a market price. The problem with either the full treatment, or the compromise treatment, is that they do not represent what bankers or borrowers do. Bankers do not buy state contingent claims from borrowers, and they do not issue price schedules relating interest charged to borrower's equity and amount lent. As indicated in my papers, bankers do set up major loan categories with different interest rates and rationing rules for each category, and my formulation of the problem was intended to minimize the violence done to these stylized facts. One of my main points was that if we think of the borrower as rationing equity size to the lender, and the lender as rationing loan size to the borrower, these quotas would have implicit prices. This defines implicitly an enriched price space which, if examined more carefully, might not look so imperfect as it does at first blush.

I do not know why bankers and other lenders do not fine-tune interest rates to the conditions of every loan. A legitimate hypothesis is that the costs of information, monitoring, negotiating, and transacting do not make it worthwhile. Rationing and qualification rules, laid down by bank management, can be applied routinely by minor officers. Also it takes a certain amount of sophistication for a borrower to understand that there are defensible reasons why his mortgage rate would be higher than his neighbor's, and bankers might like to avoid having to make explanations. So "if you want to buy a house here is the rate, and,

* University of Arizona.

providing you qualify, we will loan not more than 80 percent of market value." I am not yet prepared to say that this scenario represents either an imperfect (except by an irrelevant definition of "perfect") or inefficient market. I confess to frustration at not having a competitive process that leads to this result as an equilibrium, but I think some form of direct modeling of observed institutions is an important source of insight. Models based upon more idealized institutions can be useful if they help us to understand observed institutions, but for me they are sterile if they lead merely to the conclusion that real markets are "incomplete."

## REFERENCES

E. Baltensperger, "The Borrower-Lender Relationship, Competitive Equilibrium, and the Theory of Hedonic Prices," *Amer. Econ. Rev.*, June 1976, *66*, 401–05.

D. M. Jaffee and F. Modigliani, "A Theory and Test of Credit Rationing," *Amer. Econ. Rev.*, Dec. 1969, *59*, 850–72.

G. Stigler, "Imperfections in the Capital Market," *J. Polit. Econ.*, June 1967, *75*, 287–92.

# Labor Supply and the Payroll Tax

By C. DUNCAN MACRAE AND ELIZABETH CHASE MACRAE*

In a recent issue of this *Review*, the debate between John Brittain and Martin Feldstein regarding the incidence of the social security payroll tax focused on the effect of the payroll tax on labor supply. Both authors conclude that the tax will have an effect only if supply is wage elastic. Their analyses, however, are carried out only in terms of aggregate supply and ignore the ceiling on taxable earnings.[1]

The purpose of this note is to demonstrate that the social security payroll tax will have an effect on aggregate labor supply even if individual labor supply is wage inelastic, as long as there are individuals with earnings above the ceiling. Moreover, we will also show that the tax can be labor market neutral in the sense that it does not have an effect on aggregate labor supply, only if individual supply is wage elastic. Hence, an incidence analysis which does not take into account both the existence of the ceiling and the composition of aggregate supply can be misleading.

To see the effects of the payroll tax we employ the traditional income-leisure model of individual labor supply. The individual is assumed to supply labor so as to maximize utility from income and leisure subject to a budget constraint that specifies the tradeoff between income and leisure given the wage rate and the level of unearned income. Figure 1 shows the boundary of the individual's constraint set as the broken line $ABF$ when $Y^n$ is the level of unearned income and the (negative) slope $W$ of the line $AB$ is the wage rate. For given levels of $W$ and $Y^n$, the individual will supply a quantity of labor which corresponds to the point on the line $AB$ which lies on the highest possible indifference curve. The individual's labor supply curve, which simply shows the wage-labor combinations for some given level of unearned income ($Y^n$), with no payroll tax, appears as curve $S$ in Figure 2. Assuming

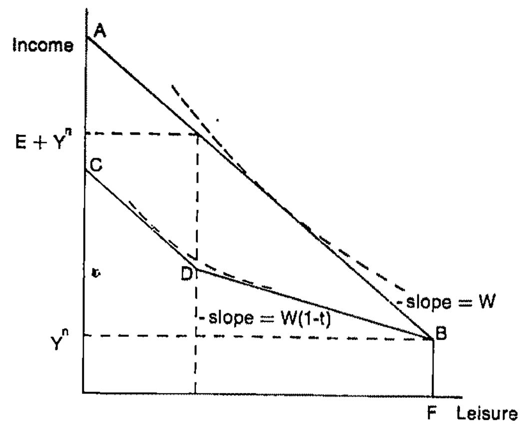



FIGURE 1

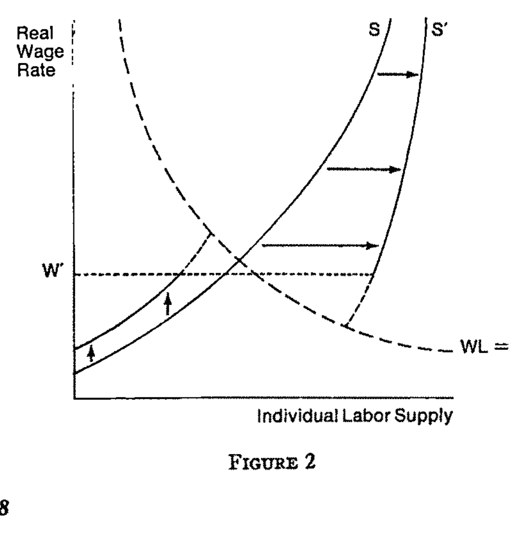



FIGURE 2

* The Urban Institute and the University of Maryland, respectively. C. D. MacRae's research was supported by funds from the Office of Research and Development, Manpower Administration, U.S. Department of Labor under Grant No. 21-11-74-09 to the Urban Institute. Opinions expressed are our own and do not necessarily represent the views of the Department of Labor, The Urban Institute or its sponsors.

[1] See Feldstein, fn. 8, p. 738, however, for a recognition that the analysis should be carried out in terms of individual supply and take the ceiling into account, although he does not do so.

that leisure is a superior good, a decrease in $Y^n$ will increase the amount of labor supplied at any wage rate and thus shift the curve to the right, perhaps with a change in shape.

Now let us introduce a payroll tax composed of a ceiling $E$ on taxable earnings and a tax rate $t$. As shown by line $CDB$ in Figure 1, if the individual does not perceive the tax as an intertemporal transfer to himself, the tax corresponds to a percentage reduction in the wage rate of $100t$ for earnings below the ceiling level $E$ and a constant reduction in income of $tE$ for earnings above $E$ with no reduction in the wage rate. The lowered slope of the budget line segment $DB$ for earnings below $E$ means that a given amount of labor will now be forthcoming only at a higher market wage rate. In Figure 2, this is shown as an upward shift in the labor supply curve in the area below the $WL=E$ hyperbola, which will decrease labor supply unless the income effect dominates. The $WL=E$ hyperbola shows the combination of wage rates and hours of labor equal to the ceiling amount of taxable earnings. The downward shift in the budget line segment $CD$ for earnings above $E$ amounts to a fixed reduction in unearned income in the amount of the tax paid, $tE$. Consequently, the labor supply curve $S'$ in Figure 2 is shown shifted to the right in the area above the $E$ hyperbola so that labor supply will definitely increase in this area if leisure is a superior good.

Since the budget line after the imposition of the payroll tax is no longer convex, there is a "jump" in the labor supply curve in the neighborhood of $E$. That is, there is some wage rate, shown as $W'$ in Figure 2, at which the individual is indifferent between supplying enough labor to generate earnings above $E$ and supplying less labor to generate earnings below $E$. Under the payroll tax, there is no combination of $W$ and $Y^n$ for which the individual will choose to earn exactly the ceiling amount $E$, since he can always increase utility by earning more or less than $E$.

The conclusion to be drawn from the foregoing analysis is that while for individuals with earnings below the ceiling the payroll tax acts to reduce their wage rate and, therefore, reduces their work incentive if supply is wage elastic, for individuals with earnings above the ceiling the tax acts solely to reduce their unearned income and, therefore, increases their work incentive if leisure is a superior good. Therefore, if individual supply is wage inelastic, the effect of the payroll tax is solely to increase the work effort of those with earnings above the ceiling and, hence, unequivocably to increase aggregate labor supply. However, if individual supply is elastic, then it is possible for the increase in supply by those above the ceiling to be cancelled out by the decrease in supply by those with earnings below ceiling. Only then will the social security payroll tax not affect aggregate labor supply, if individuals perceive it truly to be a tax and not a subsidy to themselves.

## REFERENCES

J. A. Brittain, "The Incidence of the Social Security Payroll Tax," *Amer. Econ. Rev.*, Mar. 1971, *61*, 110–25.

———, "The Incidence of the Social Security Payroll Tax: Reply," *Amer. Econ. Rev.*, Sept. 1972, *62*, 739–42.

M. Feldstein, "The Incidence of the Social Security Payroll Tax: Comment," *Amer. Econ. Rev.*, Sept. 1972, *62*, 735–38.

# On the Regulated Price-Setting Monopoly Firm with a Random Demand Curve

By Stylianos Perrakis*

The development of a substantial body of literature dealing with probabilistic microeconomics (see, for instance, Hayne Leland, John McCall, and Kenneth Smith) has imposed the necessity to reexamine certain classical results in the theory of the firm under deterministic conditions. One such category of results is derived from the analysis of the behavior of a monopoly firm under a regulatory constraint by, among others, Harvey Averch and Leland Johnson, William Baumol and Alvin Klevorick, and Akira Takayama. In view of their importance for regulatory policy these results have generated a substantial amount of controversy and discussion.[1] The question that poses itself, therefore, is to what extent is the behavior of a monopoly firm modified if the demand curve facing the monopolist is not fixed (as it was assumed until now), but fluctuates randomly according to a known probability distribution.

The general conceptual framework for the theory of the firm with uncertain demand was established recently by Leland. His study distinguishes two types of monopoly behavior, depending upon the time span of the "static" period considered and the degree of latitude allowed to the firm in order to vary the price of its output within this period. The quantity-setting firm makes its pricing decisions after the value of the random demand is known, while the price-setting firm must set the price in advance. The resulting profit-maximizing behavior is, of course, substantially different for these two classes of firms.

In an earlier study I examined the behavior of a quantity-setting monopoly firm under rate of return regulation using a fairly general set of assumptions. Focus was placed upon the well-known result of the deterministic case that the capital stock of the regulated monopolist (when the regulatory constraint is effective) is larger than the capital stock of the same firm in the absence of regulation.[2] It was shown that this result was not in general true for a monopoly firm facing a random demand curve in the absence of additional restrictions upon the shape of the probability distribution of the random factor, the monopolist's utility function, and/or the firm's revenue function. This conclusion, in turn, invalidated a number of other related results that were based on the so-called Averch-Johnson effect of overcapitalization.

In this paper I examine the behavior of a price-setting monopoly firm with random demand that is institutionally constrained to satisfy all demand for its product. As in my earlier study, one important deterministic result is not true in this more general model. This result is best described by quoting from Johnson: "Only in the unusual circumstance where the marginal physical product of capital falls below zero would the firm be led to gold-plate or to hold capital strictly idle . . . " (p. 91).[3] In fact, it is shown below under quite plausible assumptions that the regulated price-setting monopoly firm will maximize its profits by operating its capital

* Faculty of Management Sciences, University of Ottawa. I wish to thank John Zerbinis and the referee for helpful advice and comment.

[1] See, for instance, the exchange between Mohamed El Hodiri and Takayama and Israel Pressman and Arthur Carol.

[2] The most rigorous and general proof of this is in El Hodiri and Takayama. The importance of the result lies in the fact that it may be used as the basis for the proof of the so-called Averch-Johnson effect of overcapitalization (see Averch and Johnson, and Baumol and Klevorick), which is that rate of return regulation leads to a larger capital-labor ratio of the regulated firm than the one corresponding to minimum cost production at the chosen output.

[3] A proof of Johnson's statement is given by Edward Zajac.

stock at less than full capacity for an entire range of values of the random factor in the demand curve.

I shall first derive the profit-maximizing conditions for an unregulated monopolist. Next, after adding the rate of return constraint, two important propositions will be proved that define the profit-maximizing behavior of the regulated monopoly firm. Finally, the results will be extended to two special cases of random demand curves, the additive and the multiplicative random factors. In all cases the results will be proved for both expected profit and expected utility maximizing firms, hereafter referred to as risk neutrality and risk aversion, respectively.

## I. The General Model

As usual, $F(uK, L)$ denotes the monopolist's production function ($K$ capital, $L$ labor), with $F_K \equiv \partial F/\partial(uK) > 0$, $F_L \equiv \partial F/\partial L > 0$,[4] for $uK$, $L$ in a given region in the nonnegative orthant that contains the origin. The term $u$ is a linear scale variable with $0 \leq u \leq 1$, as in Smith, and Perrakis and Izzet Sahin. The usual second-order conditions are assumed to hold for $F(uK, L)$. The monopolist's demand curve is given by $D(p, x)$, where $p$ is the price and $x$ the random factor. It is assumed that $-\infty < \underline{x} \leq x \leq \bar{x} < +\infty$, and that $g(x)$, the subjective probability density function of $x$, exists and is known and continuous for $x \in X$, where $X$ denotes the closed set $[\underline{x}, \bar{x}]$. It is also assumed, following Leland, that $D_x$ is continuous and positive for all $x \in X$, and that $D_p < 0$ for all $p > 0$.

The monopolist selects the *ex post* controls $u$ and $L$ given $p$, $K$, and $x$. The *ex ante* controls $p$ and $K$ are chosen by maximizing expected utility, which is a function of profit that is linear (risk neutral) or strictly concave (risk averse). The monopolist is institutionally constrained to satisfy all demand for his product,[5] which means that:

$$F(uK, L) = D(p, x) \tag{1}$$

Given this constraint the choice of the *ex post* controls for the unregulated monopolist is made by maximizing the utility of profit $U(\pi)$ where $U' > 0$, $U'' \leq 0$, and

$$\pi(x) = pD(p, x) - wL - rK \tag{2}$$

where $w$ and $r$ are input prices. It is easy to see that the maximum utility given $p$, $K$, and $x$ is achieved by choosing the minimum $L$ satisfying (1), which is given by the choice of controls:

$$u = 1, \qquad L = \bar{L}(p, K, x) \tag{3}$$

where $\bar{L}(p, K, x) > 0$ is the (assumed unique) solution of (1) if $u = 1$.[6] The assumed restrictions on the partial derivatives of $F$ and $D$ guarantees the existence of the partials of $\bar{L}$ which are as follows:[7]

$$\bar{L}_p \equiv \frac{\partial \bar{L}}{\partial p} = \frac{D_p}{F_L}, \quad \bar{L}_x \equiv \frac{\partial \bar{L}}{\partial x} = \frac{D_x}{F_L}, \quad \bar{L}_K \equiv \frac{\partial \bar{L}}{\partial K} = -\frac{F_K}{F_L} \tag{4}$$

Hence, the *ex ante* controls $p$ and $K$ are now chosen by maximizing $E[U(\pi(x))]$ with respect to $p$ and $K$. The derivation of the necessary conditions is as follows:

$$\max_{p,K} E\{U(pD(p, x) - w\bar{L}(p, K, x) - rK)\} \tag{5}$$

$$E[U'[pD_p + D - w\bar{L}_p]] = 0 \tag{6a}$$

$$E[U'[-w\bar{L}_K - r]] = 0 \tag{6b}$$

where $U' \equiv U'(\pi(x))$, and $\pi(x)$ is given by (2) with $L = \bar{L}(p, K, x)$. If (4) is taken into account then (6a) and (6b) become:

$$E[U'D] = E\left[U'D_p\left(\frac{w}{F_L} - p\right)\right] \tag{7a}$$

[4] Hereafter subscripts will denote partial derivatives with respect to the variable(s) in the subscript.

[5] If this assumption is not introduced then the model must provide for product rationing, storage, and/or disposal.

[6] It is easy to see that the existence of an $L > 0$ satisfying (1) is guaranteed by the condition $F_L > 0 \forall L > 0$.

[7] It should be noted in (4) that the partial derivatives of $\bar{L}(p, K, x)$ are *ex ante* random variables, since $x$ enters the expressions either explicitly or implicitly through $L$.

$$E\left[U'\frac{F_K}{F_L}\right]=\frac{r}{w}E(U') \tag{7b}$$

In the case of risk neutrality $U'$ is a constant and disappears from both sides of (7a) and (7b).

Now let $(p_A, K_A)$ denote the solution of system (7a) and (7b), and assume that $(p_A, K_A)$ is unique and satisfies the second-order maximization conditions. The assumed conditions on $F$ and $D$ in turn guarantee the existence of an open neighborhood $\Omega$ of $(p_A, K_A)$ in $p$–$K$ space, in which $\bar{L}(p, K, x)$ as defined by (3) is $>0\ \forall\ x\in X$. Hereafter, it will be assumed that all pairs $p$–$K$ in our discussion fall within $\Omega$.[8]

## II. The Rate of Return Constraint

If $s>r$ is the allowable regulatory upper limit on the rate of return then we must have:

$$sK \geq pD(p, x) - wL \tag{8}$$

and the *ex post* controls are found by defining $u\in[0, 1]$ and $L>0$ satisfying (1) and (8) and maximizing $\pi(x)$ for given values of $p$, $K$, and $x$. Equation (8) may be written as:

$$L \geq \frac{pD(p, x) - sK}{w} \tag{9}$$

It is obvious now that if $\bar{L}(p, K, x)$ is $\geq$ the right-hand side of (9) then the constraint is inoperative for this particular value of $x$ and the *ex post* controls are given by (3). If, on the other hand, the right-hand side of (9) is greater than $\bar{L}(p, K, x)$, then the optimal choice is given by:

$$L=\frac{pD(p, x) - sK}{w} \tag{10}$$

$$F\left[uK, \frac{pD(p, x) - sK}{w}\right] = D(p, x)$$

the solution of the second part of (10) yielding the value of $u\in(0, 1)$. Since this value of $u$ is $<1$ it follows that for this class of values of $x$ the firm will operate *below full capacity.*

Now let $(p_B, K_B)$ be the solution of the constrained problem. Assume again that $(p_B, K_B)$ is the unique solution of the corresponding first-order necessary conditions that satisfies also the second-order conditions for a maximum. The derivation of these conditions follows from (5), with (10) replacing $\bar{L}(p, K, x)$ where appropriate. Two sets of values of $x$ may be defined as follows:

$$\omega_1 = X\cap\left\{x: sK_B \geq p_B D(p_B, x) - w\bar{L}(p_B, K_B, x)\right\} = X\cap\left\{x: F\left[K_B, \frac{p_B D(p_B, x) - sK_B}{w}\right] \leq D(p_B, x)\right\} \tag{11}$$

$$\bar{\omega}_1 = X\cap\left\{x: F\left[K_B, \frac{p_B D(p_B, x) - sK_B}{w}\right] > D(p_B, x)\right\} \tag{12}$$

and the *ex post* controls are now (3) for $x\in\omega_1$ and (10) for $x\in\bar{\omega}_1$. We may now prove the following:

PROPOSITION 1: *If the regulation is effective,*[9] $\bar{\omega}_1$ *cannot be empty for either the risk-neutral or the risk-averse monopolist.*

PROOF:

If $\bar{\omega}_1$ were empty then at $(p_B, K_B)x\in\omega_1 \Leftrightarrow x\in X$. This means that (3) is the solution of the *ex post* problem for all $x\in X$ and, hence, that (7a) and (7b) are the necessary conditions for the *ex ante* controls. However, it was assumed earlier that $(p_A, K_A)$ was the unique solution of the unconstrained problem, hence the only possibility is $(p_A, K_A) = (p_B, K_B)$. This means that the constraint (8) does not affect the choices of the monop-

[8] If $F(uK, L)$ is defined for all $(uK, L)$ in the nonnegative orthant, and if $D(p, x)$ is defined and $>0$ for all $p>0$, then $\Omega$ consists of the entire positive orthant in $p-K$ space.

[9] An effective regulation (or effective constraint) is defined here as a constraint that changes the input and/or scale choices of the monopolist for some $x\in X$.

olist for any value of $x \in X \Rightarrow$ regulation is not effective since the expected profit and/or utility are unchanged. Hence, if the regulation is effective $\bar{\omega}_1 \neq \phi$.

COROLLARY 1: *The effectively regulated price-setting monopoly firm will always operate below full capacity for some range of values of the random demand.*

As mentioned earlier, this result is drastically at variance with the deterministic results of Zajac and Baumol-Klevorick. We also note that this holding of idle capital does not happen in the unregulated monopoly with random demand, as seen by (3). Finally, nothing is implied here about the effect of the regulatory constraint on the relative size of capital stocks, prices, and/or input ratios.

The profit function of the regulated firm is now given by:

$$\text{(13a)} \quad \pi(x) = p_B D(p_B, x) - w\bar{L}(p_B, K_B, x) - rK_B, \qquad x \in \omega_1$$

$$\text{(13b)} \quad \pi(x) = (s - r)K_B, \qquad x \in \bar{\omega}_1$$

On substituting (13a) and (13b) into (5) we get:

$$\text{(14)} \quad \max_{p,K} \left\{ \int_{\omega_1} U[pD(p, x) - w\bar{L}(p, K, x) - rK]g(x)dx + \int_{\bar{\omega}_1} U[(s - r)K]g(x)dx \right\}$$

It is easy to see that the maximization procedure together with (4) yields, instead of (7a) and (7b):

$$\text{(15a)} \quad \int_{\omega_1} U'\left[pD_p + D - \frac{wD_p}{F_L}\right] g(x)dx = 0$$

$$\text{(15b)} \quad \int_{\omega_1} U'\left[\frac{wF_K}{F_L} - r\right] g(x)dx + (s - r)U' \int_{\bar{\omega}_1} g(x)dx = 0$$

where $U' \equiv U'(\pi(x))$ and $\pi(x)$ is defined by (13a) and (13b). Alternatively, (15a) and (15b) may be written as:

$$\text{(16a)} \quad \int_{\omega_1} U'Dg(x)dx = \int_{\omega_1} U'D_p\left(\frac{w}{F_L} - p\right) g(x)dx$$

$$\text{(16b)} \quad \int_{\omega_1} U' \frac{wF_K}{F_L} g(x)dx + sU' \int_{\bar{\omega}_1} g(x)dx = r\left[\int_{\omega_1} U'g(x)dx + U' \int_{\bar{\omega}_1} g(x)dx\right] = rE[U'(\pi(x))]$$

In the case of risk neutrality, (16a) and (16b) become:

$$\text{(17a)} \quad \int_{\omega_1} Dg(x)dx = \int_{\omega_1} D_P\left(\frac{w}{F_L} - p\right) g(x)dx$$

$$\text{(17b)} \quad \int_{\omega_1} \frac{wF_K}{F_L} g(x)dx + s\int_{\bar{\omega}_1} g(x)dx = r$$

The following proposition helps us characterize the solution of (16a) and (16b) and/or (17a) and (17b).

PROPOSITION 2: *$\omega_1$ possesses a relative interior for all risk-neutral and risk-averse monopolists and all distributions $g(x)$.*

PROOF:

Proposition 2 basically says that $sK_B > p_BD(p_B, x) - w\bar{L}(p_B, K_B, x)$[10] for some ranges of values of the random variable $x$. To prove it we note that the necessary conditions (16b) or (17b) are violated if $\omega_1 = \phi$, since $s$ is $> r$ by assumption. Hence we must eliminate the possibility that $\omega_1 = \{x: sK_B = p_BD(p_B, x) - w\bar{L}(p_B, K_B, x)\} \cap X$. If this is true, then $\omega_1$ is either countable or contains a compact subinterval of values of $x$ in $X$. Within such a subinterval we must have:

$$\text{(18)} \quad sK_B = p_BD(p_B, x) - w\bar{L}(p_B, K_B, x)$$

[10] Or, alternatively, that $F(K_B, (p_BD(p_B, x) - sK_B)/w) < D(p, x)$.

Differentiating (18) and taking (4) into account, we get:

$$p_B D_x = w \frac{D_x}{F_L} \tag{19}$$

and, since $D_x > 0$, we must have:

$$\frac{p_B}{w} = \frac{1}{F_L} \tag{20}$$

where

$$F_L \equiv F_L\left[K_B, \frac{p_B D(p_B, x) - sK_B}{w}\right] \tag{21}$$

Equation (20) implies that $F_L$ is constant with respect to $x$, which implies that $\partial F_L/\partial x = 0 \Rightarrow F_{LL}(p_B D_x/w) = 0$, impossible for a positive $p_B$, since $D_x > 0$ and $F_{LL} < 0$.[11] Therefore, since (18) cannot hold for a compact subinterval of values of $x$ in $X$, $\omega_1$ either contains values of $x$ in $X$, for which $sK_B > p_B D(p_B, x) - w\bar{L}(p_B, K_B, x)$, or consists entirely of a countable set of single points, on which (18) holds. The latter case would violate the necessary conditions (15b) or (17b), since the first integral on the left-hand side would be zero because of the continuity of $g(x)$. The only remaining possibility is that $sK_B > p_B D(p_B, x) - w\bar{L}(p_B, K_B, x)$ for some $x$ in $X$ and the continuity of $D$ and $\bar{L}$ insures that the above inequality will hold strictly for a range of values of $x$ in $X$.

Propositions 1 and 2 imply that there exist at least two compact intervals of values of $x$ in $X$, in which constraint (8) holds with equality and with strict inequality respectively. The consequences of this may be expressed by the following:

COROLLARY 2: *The expected rate of return on invested capital for the regulated price-setting monopolist is strictly less than the regulatory upper bound s.*[12]

[11] For a neoclassical production function with convex isoquants $F_{LL}$ must be $<0$.

[12] Corollary 2 is also valid for the quantity-setting regulated firm, as can be seen from my earlier study. This is a consequence of the regulatory assumptions of both studies, which imply that the regulators would enforce the regulatory constraint at every observable state of nature, and the firm knows this and acts accordingly. A different model, requiring a multiperiod formulation would be more appropriate if it is assumed that the regulators restrict the *average* rate of return over a number of observed realizations of the random demand.

## III. Some Special Cases

Two important special forms of the random demand curve are those with additive and multiplicative random terms. Specifically:

$$D(p, x) = D(p) + x > 0 \quad \text{for additive} \tag{22a}$$

$$D(p, x) = xD(p), \quad x > 0 \quad \text{for multiplicative} \tag{22b}$$

We may assume without loss of generality that $E(x) = 0$ in (22a) and $E(x) = 1$ in (22b).

The solution of the problem of the constrained price-setting monopolist involves, among others, the definition of the subspaces $\omega_1$ and $\bar{\omega}_1$. For this we need to find the roots in $x$ of the equation

$$\frac{F[K, pD(p, x) - sK]}{w} = D(p, x) \tag{23}$$

This is simplified in both additive and multiplicative cases by the following:

PROPOSITION 3: *For random demand curves with additive or multiplicative random factors $\omega_1$ is compact and* (23) *has either one or two roots in X.*

PROOF: Define $\Lambda(x)$ as follows:

$$\Lambda(x) = \frac{F[K, pD(p, x) - sK]}{w} - D(p, x) \tag{24}$$

It is easy to see by repeated differentiation that $\Lambda''(x) < 0$ when $D(p, x)$ has the forms (22a) or (22b). Hence, $\Lambda(x)$ is strictly concave and the equation $\Lambda(x) = 0$ has at most two roots in $X$. Since $\bar{\omega}_1$ is defined by $\Lambda(x) > 0$ the strict concavity of $\Lambda(x)$ insures that if $\hat{x}_1 < \hat{x}_2$ are the two roots of (23), then $\bar{\omega}_1 = (\hat{x}_1, \hat{x}_2) \cap X$, which is compact.

The fact that $\omega_1$ must possess an interior implies that at least one of the subspaces $[\underline{x}, \hat{x}_1] \cap X$, $[\hat{x}_2, \bar{x}] \cap X$ must possess an interior. It is easy to show that it is not

possible to have $\omega_1 = [\hat{x}_2, \bar{x}] \cap X$. To see this we observe that at $(p_B, K_B)$ the point $x_m$, at which $\Lambda(x) = max$ is given by the solution of $\Lambda'(x) = 0$, which yields $pF_L/w = 1$; obviously, if $x_m \in X$ then $x_m \in \bar{\omega}_1$. It follows that $\hat{x}_2 > x_m$ and, if $\omega_1 = [\hat{x}_2, \bar{x}]$, $\Lambda'(x) < 0 \ \forall \, x \in \omega_1$, implying in turn that $w/F_L - p > 0 \ \forall \, x \in \omega_1$ and the necessary condition (16a) can have no solution since the left-hand side is $>0$ and the right-hand side $<0$ always. Hence, for these two important special forms of the random demand curve we must have either:

$$\omega_1 = [\underline{x}, \hat{x}_1], \qquad \bar{\omega}_1 = (\hat{x}_1, \bar{x}) \tag{25}$$

or

$$\omega_1 = [\underline{x}, \hat{x}_1] \cup [\hat{x}_2, \bar{x}], \quad \bar{\omega}_1 = (\hat{x}_1, \hat{x}_2) \tag{26}$$

where $\hat{x}_1$ and $\hat{x}_2$ are the two roots in $x$ of (23) for the optimal *ex ante* controls $(p_B, K_B)$.

## IV. Conclusions

The combination of demand uncertainty, price-setting behavior, and rate of return regulation for a monopoly firm has as a result the operation of the firm's capital stock below full capacity for at least one entire range of realizations of the random demand at the optimally chosen price and capital. This is true for all types of probability distributions and all utility functions with nondecreasing marginal utility, which includes of course, risk-neutral monopolists. Further, the expected *ex ante* rate of return is strictly less than the maximum allowed by the regulators, while the observed *ex post* rate of return is equal to the maximum allowed whenever the firm operates its capital stock below full capacity, and vice versa.

These results are direct consequences of the adopted assumptions about firm objectives and regulation, which are natural extensions to the uncertain demand case of the Averch-Johnson model of firm behavior. The firm must satisfy *ex post* (i.e., after the value of the random demand becomes known) both the rate of return constraint and the demand constraint (1). At some "states of nature," it must increase its labor above a certain minimum in order to stay within the allowed rate of return. This, however, forces it to underutilize the invested capital in order to meet demand exactly. These states of nature will occur with a positive probability whenever the firm choooses *ex ante* the output price and capital stock by maximizing expected profit (or expected utility of profit). Needless to say, the conclusions of this paper will have to be modified if other firm objectives or methods of regulation are considered (as in the article by Elizabeth Bailey and John Malone for the deterministic case).

As a final remark it should be noted that, although numerical solution of (23) and (16a) and (16b) should present no problems, analytical results are difficult to obtain. Proposition 3 offers some help in this direction for the two important cases of demand curves with additive or multiplicative random terms, since it shows that the solution of (23) is equivalent to finding the zeros of a strictly concave function. However, even in the simple case of a Cobb-Douglas production function and a constant elasticity demand curve with a multiplicative random factor it is not possible to solve (23) explicitly. Barring restrictive assumptions, therefore, there is very little that can be said about relative size of capital stock, price, expected output, etc.

## REFERENCES


**H. Averch and L. Johnson**, "Behavior of the Firm under Regulatory Constraints," *Amer. Econ. Rev.*, Dec. 1962, *52*, 1052–69.

**E. Bailey and J. Malone**, "Resource Allocation and the Regulated Firm," *Bell J. Econ.*, Spring 1970, *1*, 129–42.

**W. Baumol and A. Klevorick**, "Input Choices and Rate-of-Return Regulation: An Overview of the Discussion," *Bell J. Econ.*, Fall 1970, *1*, 162–90.

**M. El-Hodiri and A. Takayama**, "Behavior of the Firm Under Regulatory Constraint: Clarifications," *Amer. Econ. Rev.*, Mar. 1973, *63*, 235–37.

**L. L. Johnson**, "Behavior of the Firm Under Regulatory Constraint: A Reassessment," *Amer. Econ. Rev. Proc.*, May 1973, *63*, 90–97.

H. Leland, "Theory of the Firm Facing Uncertain Demand," *Amer. Econ. Rev.*, June 1972, *62*, 278–91.

J. J. McCall, "Probabilistic Microeconomics," *Bell J. Econ.*, Fall 1971, *2*, 403–33.

S. Perrakis, "Rate-of-Return Regulation of a Monopoly Firm with Random Demand," presented at the December 1973 meetings of the Econometric Society (rev., *Int. Econ. Rev.*, forthcoming).

——— and I. Sahin, "Resource Allocation and Scale of Operations in a Monopoly Firm: A Dynamic Analysis," *Int. Econ. Rev.*, June 1972, *13*, 399–407.

I. Pressman and A. Carol, "Behavior of the Firm Under Regulatory Constraint: Note," *Amer. Econ. Rev.*, Mar. 1971, *61*, 210-12.

——— and ———, "Behavior of the Firm Under Regulatory Constraint: Reply," *Amer. Econ. Rev.*, Mar. 1973, *63*, 238.

K. R. Smith, "The Effect of Uncertainty on Monopoly Price, Capital Stock and Utilization of Capital," *J. Econ. Theory*, June 1969, *1*, 48–59.

A. Takayama, "Behavior of the Firm Under Regulatory Constraint," *Amer. Econ. Rev.*, June 1969, *59*, 255–60.

E. E. Zajac, "Note on 'Gold Plating' or 'Rate Base Padding'," *Bell J. Econ.*, Spring 1972, *3*, 311–15.

# Money, Income, and Causality: The U.K. Experience

*By* DAVID WILLIAMS, C. A. E. GOODHART, AND D. H. GOWLAND*

In a recent paper, Christopher Sims developed a statistical technique for testing whether there was evidence of unidirectional causality within a two-variable system. He applied this technique to an examination of the relationship between movements in nominal incomes and in the money supply, in order to throw light on the question whether there was evidence of clear-cut unidirectional causality running from changes in the money stock to changes in nominal incomes, or vice versa. He concludes that: "The main empirical finding is that the hypothesis that causality is unidirectional from money to income agrees with the postwar *U.S.* data, whereas the hypothesis that causality is unidirectional from income to money is rejected" (p. 540). Accordingly,

> . . . one clearly should not estimate a demand for money relation from these data [of *GNP* and money stock], treating *GNP* as exogenous with money on the left-hand side; no evidence appears to contradict the common assumption that money can be treated as exogenous in a regression of *GNP* on current and past money. [p. 550]

This is an important, indeed sweeping, conclusion, which could if generally upheld have a major influence on monetary economics. One way of examining the strength of these results is to replicate the exercise in other contexts and using differing data and time periods. In this paper we undertake a closely similar exercise using *U.K.* data.

There were a number of reasons for doubting whether Sims' results would hold within a *U.K.* context. Unlike the United States, the United Kingdom is a relatively small, open economy which during the observation period maintained a regime of fixed exchange rates. Under such circumstances the private sector can more easily adjust their money holdings to their incomes by transfers of funds over the exchanges (subject to exchange control), than in a virtually closed economy such as the *U.S.* Furthermore the conduct of monetary policy during the 1960's was influenced by the precepts of the Radcliffe Committee Report that "The authorities thus have to regard the structure of interest rates rather than the supply of money as the centre-piece of the monetary mechanism" (p. 135). In so far as the authorities primarily aim to regulate the structure of interest rates, movements in the money stock can be expected to respond to movements in nominal incomes.

Because of the various differences in context the finding that in the United Kingdom the relationship between money and income appears different from that found by Sims for the United States in no way casts any doubt on the validity of Sims' own results. Sims, in fact, comments (pp. 542–43) that the results are peculiar to his sample. A recent study by Thomas Sargent and Neil Wallace[1] investigates the direction of causality between money and prices during periods of hyperinflation for certain European countries using an approach similar to Sims. They show that there is evidence to suggest that the causality is from prices to money. Hence, the implication is that the form and direction of these relationships do depend on the institutional context, and that Sims' results do not have a general validity in all economies. In particular it may be hazardous to extrapolate findings obtained in a virtually closed economy such as the *U.S.*, also to smaller, open economies.

* Bank of England. Our thanks are due to our colleagues in the Economic Section of the Bank for their general help and to a referee of an earlier version of this paper for some specific advice and suggestions. In addition, we have benefited by having the chance to read as yet unpublished papers on this topic by E. L. Feige of the University of Wisconsin and K. Wall of Queen Mary College, London. The errors and shortcomings, however, remain our responsibility.

[1] We are grateful to a referee for bringing this reference to our attention.

## I. Methodology and Content

Following an approach initially developed by C. W. J. Granger Sims sets out and proves the theorem that when the jointly covariance-stationary pair of stochastic processes $X$ and $Y$ . . .

> . . . has an autoregressive representation, $Y$ can be expressed as a distributed lag function of current and past $X$ with a residual which is not correlated with any values of $X$, past or future, if, and only if, $Y$ does not cause $X$ in Granger's sense. . . . Only in the special case where causality runs from $X$ to $Y$ can we expect that no future values of $X$ would enter the regression if we allowed them. [p. 544–45]

So if $X$ causes $Y$, a regression of $Y$ on past, current, and future values of $X$ should exhibit significant coefficients for past and current values but insignificant coefficients on future values. Again if $X$ causes $Y$, a regression of $X$ on past, current, and future values of $Y$ should exhibit significant coefficients on future values of $Y$ (and may, or may not, exhibit significant coefficients on present and past values). As Sims states, this method "does rest on a sophisticated version of the *post hoc ergo propter hoc* principle," but he argues strongly that this "method is not easily fooled" (p. 543).

Essentially, therefore, the method involves regressing $X$ on current, past, and *future* values of $Y$, and then reversing the process regressing $Y$ on $X$. Causality, in this sense, then appears to be unidirectional from $X$ to $Y$, if future values of $X$ are insignificant, but current and past values of $X$ are significant in the regression of $Y$ on $X$, while future values of $Y$ are significant when regressing $X$ on $Y$. These tests will only indicate the presence of what may be called proper causality, i.e., the causal effect takes at least one time period to manifest itself. Contemporaneous causality can be tested by including the present with the future lags when carrying out the tests. However, we followed Sims in only testing for proper causality.

Sims noted (p. 543) that there are two situations when the results of these tests can be misleading:

(i) The first situation concerns the possible existence of a third variable which, say, affects $X$ quicker than $Y$. In this case the future lags of $Y$ will be significant when explaining $X$ even though $X$ and $Y$ themselves may not be causally related. We will never know whether this situation has obtained without formulating a more comprehensive model. This is something of a vicious circle since these causality tests are to be used as a preliminary to any model building.

(ii) The second situation arises when the factors affecting one variable, say $Y$, include the expected future values of the other variable $X$ and yet there is no causal relationship from $Y$ to $X$. In the event of the expectations being accurate, the future values of $X$ may be significant when explaining $Y$ owing to the autocorrelation of $X$. This situation should rarely occur in practice because of the assumption that there is no causal relationship from $Y$ to $X$. However Michael Rozeff, in a recent study analyzing the relationship between money and stock market prices in the United States, has suggested that investors do correctly anticipate the future path of monetary expansion although there is supposedly no causal relationship from stock market prices to money.

For this exercise, examining the direction of causality between incomes and money, the standard *U.K.* series for nominal incomes is the quarterly series for *GDP* at current prices. This series can be divided into its real (*RGDP*) and price components (*PGDP*), which are additive on the logarithmic scale. This series is available on a consistent basis from 1958-I till 1971-III, the period of observation for this exercise. Complete monetary data, covering all the banks, are only available from 1963-I onwards, so, in order to obtain a longer period of observation, proxy series were used. Two monetary series were used: (i) notes and coin plus current accounts with the London clearing banks, subsequently termed narrow money ($M_N$); (ii) $M_N$, plus deposit accounts with the London clearing banks, subsequently termed broad money ($M_B$).

The first may be regarded as a proxy for $M_1$, and the second for $M_3$ as currently defined (Bank of England). These monetary aggregates based on London clearing bank deposits formed a declining proportion of the total money stock, especially in the case of the broader definition in the later years. However, the London clearing banks account for a very large proportion of the *U.K.* banking sector, and throughout the period maintained a dominant share of current accounts, over 80 percent. Therefore, $M_N$ is probably a better proxy for $M_1$, than is $M_B$ for $M_3$. It should, however, also be noted that the clearing banks tended to bear the brunt of the authorities' direct interventions in the banking system (for example, ceilings on lending). These monetary series were available monthly and were converted into a consistent quarterly series by averaging. Since these variables are only proxies, then, only to a limited extent are they measuring the theoretical concepts which they are representing. This is typically an instance of errors in variables. These errors will tend to reduce the significance of the coefficients in the Sims regressions and so causality is less likely to be detected, and might be distorted. However, there is no reason to expect these measurement errors in money to be correlated with incomes, unlike the case cited by Sims (p. 543) where "animal spirits" of bankers and businessmen could affect money quickly but *GNP* only with a distributed lag.

Sims noted (p. 546) that if different seasonal adjustment procedures are used on money and incomes then "spurious" seasonality can be induced into the relationship. We used the same seasonal adjustment procedure on both sides of the regression equation, and could therefore dispense with seasonal dummies. However, we tested the residuals from each fitted regression for fourth-order autocorrelation and found it to be absent in every case (see below).

Before computing his regressions, Sims applied a filter to all his variables of the form $(1-\frac{3}{4}B)^2$ where $BX_t = X_{t-1}$. The object of this filter was to ensure that the residuals were approximately white noise. Marc Nerlove considered filters of the type $(1-kB)^p$ applied to seasonally adjusted series. His objective was to remove the trend, and he judged this successful if the resulting spectrum was relatively flat, thereby simulating white noise. He found empirically that flat spectra could be achieved in quite a number of economic time-series with $k=\frac{3}{4}$ and $p=1$, 2, or 3.

As noted by Sims (p. 544), the derivation of the regression tests between two economic processes $X$ and $Y$ requires that these series are jointly covariance stationary processes each with zero mean. Since they are growing through time, both the money and income series are nonstationary. This is true for both the *U.S.* and the *U.K.* Sims assumes that the nonstationary behavior can be modelled by a deterministic linear time trend plus a stationary stochastic residual. We preferred the more general assumption of stochastic nonstationarity.[2] G. E. P. Box and G. M. Jenkins (p. 85) suggest that such nonstationarity can be eliminated by an appropriate degree of differencing, i.e., using a filter $(1-\mathrm{B})^d$ with $d$ usually equal to 1 or 2.

Like Sims, we assumed that the relationship, if any, between money and income was linear on the logarithmic scale. Using the methodology of Box and Jenkins the transformed series of both money and income can be made stationary by taking first differences. To ensure a zero mean, each stationary series should be mean corrected by subtracting the appropriate sample mean. This is equivalent to including a constant term in the regressions, which is what we did.

In specifying his regressions, Sims used 4 future and 8 past lags and also included a time trend. Before estimation he used the filter $(1-\frac{3}{4}B)^2$. Our specification can be thought of as being the same as Sims' except that we used the filter $(1-B)$ which reduces the time trend to a constant. However the

[2] This means that the parameters of the linear time trend are assumed to be stochastic. However, the data were such that only the intercept parameter could be considered stochastic. It will be seen below that, following a referee's suggestion, the assumption of stochastic nonstationarity was tested and justified empirically.

purpose of Sims' filter was to produce an uncorrelated error structure in each of the regressions whereas our filter was to produce covariance stationary processes. Hence, we possibly need a further filter to cope with autocorrelation in the error structure. Autocorrelation may be present because the chosen lag length is arbitrary and so further relevant lags are included in the error term. In order to assess the magnitude of this autocorrelation, we have made the standard assumption that it can be modelled by an autoregressive process, since this can be dealt with using regression techniques. So, instead of only applying a single filter to each regression before estimating the lag profile, we adopted the following procedures:

(i) apply *OLS* to the first-differenced specification;

(ii) investigate the autoregressive properties of the fitted residuals, $e_t$, from step (i) by regressing $e_t$ on $e_{t-1}, \ldots, e_{t-4}$. We found that $e_{t-1}$ and $e_{t-2}$ at most had significant coefficients giving a filter of the form $(1-a_1B-a_2B^2)$. The lagged error $e_{t-4}$ was used to check for residual seasonality and was found to be insignificant in every case;

(iii) apply the estimated filter from step (ii) to the first-differenced specification and, using *OLS*, estimate the lag profile.

Ideally, this procedure should be iterated using the fitted residuals from step (iii) and repeating steps (ii) and (iii). However, since it is necessary to use the same autoregressive filter in estimating the lag profile both with and without the future lags, and this common filter is estimated by averaging the two independently estimated filters, one iteration was considered sufficient. With first-differencing and the autoregressive filter, the combined filter which is applied to each variable is of the form $(1-B)(1-a_1B-a_2B^2)$ for the more complicated second-order autoregressive filter. This may be compared and contrasted with Sims' filter $(1-\frac{3}{4}B)^2$.

In our approach, the filter $(1-B)$ is assumed to be necessary. A referee suggested that it would be possible to check this assumption empirically as follows:

(i) form the appropriate specification between the unfiltered variables, i.e., with or without the future lags, and also include a time trend in the specification;

(ii) apply an initial filter to the specification, for example, either $(1-B)$ or $(1-\frac{3}{4}B)^2$, and estimate the lag profile using *OLS*;

(iii) using the estimated coefficients from step (ii), evaluate the residuals in the original specification. If the filter $(1-B)$ has been used then these residuals will not have zero mean;

(iv) investigate the autoregressive properties of these residuals $e_t$ by regressing $e_t$ on $e_{t-1}, \ldots, e_{t-4}$ and including a constant term if necessary;

(v) use the estimated filter from step (iv) instead of the initial filter and repeat step (ii).

We extended the referee's suggestion by iterating the procedure and then averaged the two final estimated filters for the specification with and without the future lags to derive a common filter which is therefore equivalent to the combined filter mentioned above. A comparison of the filters using the two approaches is shown in Table 1. With the referee's approach, four of the filters have a factor $(1-B)$ (assessed by the algebraic sum of the coefficients being zero) and half of the remaining filters have a factor close to $(1-B)$ (assessed by algebraic sum of the coefficients being less than 0.2). This suggests that our assumption of $(1-B)$ is justified. In fact, the regression results of the two approaches were similar and the conclusions identical. Consequently only the results of our approach are reported since we believe that our combined filter having a factor of $(1-B)$ is theoretically more satisfying.

## II. Empirical Results

The crucial empirical issue is whether future values appear significant when each series is regressed on a distributed lag of past, present, and future values of the other series. Sims found "... Future values of *GNP* were highly significant in explaining the *M* dependent variable, but future values of *M* were not significant in explaining the

TABLE 1—COMPARISON OF COMBINED FILTERS

| Regression | Our Approach[a] | Referee's Approach[a] |
|---|---|---|
| **Nominal Income** | | |
| $M_N$ on $GDP$ | $1-1.3B+0.6B^2-0.3B^3$ | $1-1.0B+0.45B^2$ |
| $M_B$ on $GDP$ | $1-1.4B+0.4B^2$ | $1-1.4B+0.6B^2$ |
| $GDP$ on $M_N$ | $1-0.7B-0.3B^2$ | $1-0.65B-0.35B^2$ |
| $GDP$ on $M_B$ | $1-0.75B-0.25B^2$ | $1-B$ |
| **Real Income** | | |
| $M_N$ on $RGDP$ | $1-1.4B+0.7B^2-0.3B^2$ | $1-1.0B+0.6B^2$ |
| $M_B$ on $RGDP$ | $1-1.3B+0.3B^2$ | $1-1.3B+0.5B^2$ |
| $RGDP$ on $M_N$ | $1-0.7B-0.3B^2$ | $1-0.55B-0.35B^2$ |
| $RGDP$ on $M_B$ | $1-0.7B-0.3B^2$ | $1-0.55B-0.2B^2$ |
| **Price Component** | | |
| $M_N$ on $PGDP$ | $1-1.45B+0.65B^2-0.2B^3$ | $1-1.25B+0.55B^2$ |
| $M_B$ on $PGDP$ | $1-0.7B-0.3B^2$ | $1-1.4B+0.55B^2$ |
| $PGDP$ on $M_N$ | $1-0.7B-0.5B^2+0.2B^3$ | $1-0.45B-0.55B^2$ |
| $PGDP$ on $M_B$ | $1-0.8B-0.35B^2+0.15B^3$ | $1-0.55B-0.45B^2$ |

[a] See text.

$GNP$ dependent variable" (p. 546). An $F$-test was used to check the significance of the future values. Although not as significant, our results using $U.K.$ data suggested the opposite relationship particularly with the narrow definition of money $M_N$. The situation was less clear-cut with the broad definition of money $M_B$. In Table 2, our results and those presented by Sims (in his Table 3) are shown and contrasted. None of our $F$-ratios are significant at the 5 percent level.[3] However, the future lags for $GDP$ on $M_N$ are almost significant at the 5 percent level whereas those for $M_N$ on $GDP$ are definitely nonsignificant. This suggests at least the possibility of unidirectional causality from $GDP$ to $M_N$.

In Table 3, we present the $R^2$ for the various regressions, which show that there is still a sizeable proportion of unexplained variance. A test of whether or not $R^2$ is significantly different from zero is assessed by an $F$-ratio. These are also given in Table 3. Only two ratios are significant at the 5 percent level, viz., $M_N$ on $GDP$ without future lags and $GDP$ on $M_N$ with future lags. These are the two ratios that would a priori have

[3] The significance level quoted in all the tables is for the case of a second-order autoregressive filter since this gives the minimum degrees of freedom and hence the maximum value of the $F$-ratio.

TABLE 2—$F$-TESTS ON FOUR FUTURE QUARTERS' COEFFICIENTS

| Regression | $F$-ratio |
|---|---|
| **Sims' $U.S.$ results** | |
| $GNP$ on $M_N$ | 0.36 |
| $GNP$ on $M_B$ (monetary base) | 0.39 |
| $M_N$ on $GNP$ | 4.29[a] |
| $M_B$ on $GNP$ | 5.89[a] |
| **$U.K.$ results** | |
| $GDP$ on $M_N$ (narrow money) | 2.44 |
| $GDP$ on $M_B$ (broad money) | 0.84 |
| $M_N$ on $GDP$ | 0.97 |
| $M_N$ on $GDP$ | 1.85 |
| 5 percent significance level | 2.74 |

[a] Significant at 0.05 level.

TABLE 3

| | With Future Lags | | Without Future Lags | |
|---|---|---|---|---|
| Regression | $R^2$ | $F$-ratio | $R^2$ | $F$-ratio |
| $GDP$ on $M_N$ | 0.51 | 2.20 | 0.34 | 1.77 |
| $GDP$ on $M_B$ | 0.38 | 1.26 | 0.30 | 1.48 |
| $M_N$ on $GDP$ | 0.48 | 1.82 | 0.40 | 2.21 |
| $M_B$ on $GDP$ | 0.32 | 1.00 | 0.14 | 0.56 |
| 5 percent significance level | | 2.12 | | 2.21 |

TABLE 4—ESTIMATED LAG PROFILES

| | *GDP* on $M_N$ | | $M_N$ on *GDP* | | *GDP* on $M_B$ | | $M_B$ on *GDP* | |
|---|---|---|---|---|---|---|---|---|
| Lag | | | | | | | | |
| −4 | −0.05 | * | 0.07 | * | 0.16 | * | 0.19 | * |
| −3 | 0.14 | * | 0.16 | * | −0.15 | * | −0.15 | * |
| −2 | −0.08 | * | −0.03 | * | 0.15 | * | −0.33 | * |
| −1 | 0.44 | * | 0.20 | * | 0.25 | * | −0.01 | * |
| 0 | −0.10 | 0.20 | 0.40 | 0.29 | −0.05 | 0.18 | 0.09 | 0.12 |
| 1 | 0.48 | 0.29 | 0.39 | 0.47 | −0.11 | −0.06 | 0.18 | 0.21 |
| 2 | −0.45 | −0.39 | 0.22 | 0.25 | −0.40 | −0.45 | 0.30 | 0.34 |
| 3 | 0.57 | 0.51 | −0.15 | −0.13 | 0.34 | 0.25 | 0.10 | 0.16 |
| 4 | −0.14 | −0.15 | −0.50 | −0.56 | −0.40 | −0.41 | 0.12 | 0.07 |
| 5 | 0.26 | 0.24 | −0.24 | −0.34 | 0.88 | 0.79 | −0.17 | −0.05 |
| 6 | −0.33 | −0.38 | 0.00 | 0.01 | −1.14 | −0.96 | −0.10 | −0.02 |
| 7 | 0.32 | 0.25 | −0.05 | −0.20 | 0.91 | 0.58 | 0.08 | −0.04 |
| 8 | 0.06 | 0.11 | 0.12 | 0.03 | −0.48 | −0.32 | 0.03 | −0.03 |
| Standard Error | | | | | | | | |
| max. | 0.26 | 0.27 | 0.23 | 0.21 | 0.38 | 0.35 | 0.21 | 0.21 |
| min. | 0.14 | 0.14 | 0.16 | 0.17 | 0.20 | 0.24 | 0.13 | 0.15 |

been expected to be significant on the assumption of unidirectional causality from *GDP* to $M_N$. The coefficients of the regressions themselves are given in Table 4. Caution should be exercised in interpreting both the sizes and patterns of the coefficients since unrestricted estimation is unlikely to give an efficient estimate of the lag profile.

Sims then proceeded to examine the relationship between money and the real and price components of income separately. He found that the relationship between money and the real component of *GNP* (*RGNP*) was much the same as the relationship between *M* and nominal income, but, with the price component of *GNP* (*PGNP*), none of the *F*-ratios testing the future coefficients were significant. We undertook a similar exercise and the *F*-ratios testing the future coefficients are presented in Table 5. None of these *F*-ratios for *RGDP* is significant. However with *PGDP*, that with $M_N$ as the dependent variable is almost significant at the 5 percent level. This suggests a possibility that there might be unidirectional causality from money to prices.

Again, we have estimated the significance of the $R^2$ in the various regressions. These results are presented in Table 6. None of them is significant at the 5 percent level.

TABLE 5—*F*-TESTS ON FOUR FUTURE QUARTERS' COEFFICIENTS

| Regression | *F*-ratio |
|---|---|
| $M_N$ on *PGDP* | 2.73 |
| $M_B$ on *PGDP* | 2.07 |
| *PGDP* on $M_N$ | 1.26 |
| *PGDP* on $M_B$ | 0.56 |
| $M_N$ on *RGDP* | 0.45 |
| $M_B$ on *RGDP* | 0.54 |
| *RGDP* on $M_N$ | 1.03 |
| *RGDP* on $M_B$ | 0.21 |
| 5 percent significance level | 2.74 |

TABLE 6

| | With Future Lags | | Without Future Lags | |
|---|---|---|---|---|
| Regression | $R^2$ | *F*-ratio | $R^2$ | *F*-ratio |
| *RGDP* on $M_N$ | 0.41 | 1.41 | 0.31 | 1.57 |
| *RGDP* on $M_B$ | 0.27 | 0.76 | 0.24 | 1.12 |
| $M_N$ on *RGDP* | 0.32 | 0.94 | 0.27 | 1.25 |
| $M_B$ on *RGDP* | 0.19 | 0.48 | 0.12 | 0.47 |
| *PGDP* on $M_N$ | 0.34 | 1.04 | 0.21 | 0.90 |
| *PGDP* on $M_B$ | 0.39 | 1.29 | 0.34 | 1.72 |
| $M_N$ on *PGDP* | 0.49 | 1.92 | 0.28 | 1.26 |
| $M_B$ on *PGDP* | 0.32 | 0.99 | 0.11 | 0.44 |
| 5 percent significance level | | 2.12 | | 2.21 |

This is further evidence that the relationship between movements in money and in prices and real incomes in the *U.K.* is not very close.

### III. Conclusion

The evidence indicating the direction of causality between money and income in the *U.K.* is much less clear-cut than that which Sims found in his examination of *U.S.* data. As remarked above, the tests carried out by both Sims and ourselves are for proper causality, i.e., with a lag of at least one period. A negative result still leaves open the possibility of simultaneous causality. Sims found evidence of unidirectional causality running from money to nominal income, especially from money to real income. We found for the *U.K.* some evidence of unidirectional causality running from nominal incomes to money but also some evidence of unidirectional causality running from money to prices. Taken together, this evidence suggests, perhaps, a more complicated causal relationship between money and incomes in which both are determined simultaneously.

## REFERENCES

**G. E. P. Box and G. M. Jenkins**, *Time Series Analysis, Forecasting and Control*, San Francisco 1970.

**C. W. J. Granger**, "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods," *Econometrica*, July 1969, *37*, 424–38.

**M. Nerlove**, "Spectral Analysis of Seasonal Adjustment Procedures," *Econometrica*, July 1964, *32*, 241–86.

**M. S. Rozeff**, "Money and Stock Prices; Market Efficiency and the Lag in Effect of Monetary Policy," *J. Finance.*, Sept. 1974, *1*, 245–302.

**T. J. Sargent and N. Wallace**, "Rational Expectations and the Dynamics of Hyperinflation," *Int. Econ. Rev.*, June 1973, *14*, 328–50.

**C. A. Sims**, "Money, Income, and Causality," *Amer. Econ. Rev.*, Sept. 1972, *62*, 540–52.

**Bank of England**, "Changes in Banking Statistics," *Bank of England Quart. Bull.*, Mar. 1972, *12*, 76–79.

**Radcliffe Report**, *Committee on the Working of the Monetary System*, Cmnd. 827, London 1959.

# The Neoclassical Production Function: Comment

By P. GAREGNANI*

In an article recently published in this *Review*, Lowell Galloway and Vishwa Shukla (hereafter G-S) consider an economy where two commodities, 1 and 2, are produced in separate industries and in annual cycles by means of labor and a circulating capital made up of the commodities themselves. Two alternative methods are assumed for the production of commodity 2 and there are, therefore, two alternative "systems" of production, each formed by taking the method of production of commodity 1 together with one of the methods for commodity 2; moreover, both commodities are basic in each of the systems.[1]

## I. A Nonreswitching Theorem

Galloway and Shukla claim that if each system gives a relative price of the two commodities which is "positive and finite for any positive value of the interest rate" (p. 349), then reswitching between them cannot occur; i.e., the system which has become less profitable than the other because of a change in the rate of interest $r$ cannot revert to being more profitable as $r$ changes further in the same direction.

There are two points to be made about this claim. The first is that the claim is false: under the conditions stated by G-S reswitching is possible. A numerical example showing this is given in the Appendix together with an indication of the error in their attempted proof.

The second point is that, contrary to what readers of the article might have been led to think, the G-S condition of a positive price "for any positive value of the interest rate" has no apparent economic justification.

## II. The Maximum Rate of Interest

In their article, G-S refer to numerical examples which various authors, myself included, had given in order to disprove an earlier nonreswitching "theorem,"[2] and write "... in all fairness to the counterexamples it should be noted that $p_1$ and $p_2$ [the prices of the two commodities] are positive within the range of $r$ encompassing the switch points. However at other values of $r$, either $p_1$ or $p_2$ is negative" (p. 350, fn. 10). They, however, do not specify that the "other values of $r$," at which the relative price of the two commodities[3] is negative in the counterexamples, are values of $r$ larger than the "maximum rate of interest," i.e., are rates of interest for which the wage and the relative price of the two commodities cannot both be positive.[4]

When, therefore, G-S, later in the text, write of the "Failure of ... counterexamples of reswitching to meet the basic assumption of positive and finite commodity prices for all $r>0$ ..." (p. 351), the reader may be induced to think that the examples in question are in contradiction to the usual assumption of positive prices (for rates of in-

* Professor of economics, University of Rome.

[1] The G-S assumption of "indecomposable production systems" (p. 349) is equivalent to assuming that each of the two commodities is basic, i.e., enters directly or indirectly in the production of both itself and the other commodity (see Piero Sraffa, p. 8).

[2] The nonreswitching theorem was given by David Levhari, and the counterexamples by Michio Morishima; Michael Bruno, Edwin Burmeister, and Eytan Sheshinsky; the author (1966).

[3] The price equations can only determine relative prices, and therefore, in a two-commodity system, they determine a single relative price.

[4] In any system of production which is "viable," (i.e., capable of yielding a surplus over the replacement of the means of production) and includes only basic commodities, the rise from zero of the rate of interest $r$ makes the wage in terms of any commodity fall continuously from a value $w_{max}$ for $r=0$, down to the level zero, reached for a finite "maximum rate of interest" $r_{max}$, at which the whole national income goes to profits. Further, for $0 \leq r \leq r_{max}$ all prices are positive. These properties, well known since Sraffa, are ignored by G-S, though they were restated by myself (1966) in the article to which they refer. The additional proposition that the wage and prices cannot all be positive for $r>r_{max}$ is demonstrated in paragraph 3 of the Appendix below.

terest up to the maximum) "basic" to all economic theory, and not in contradiction solely to the special assumption made by G-S of positive prices even beyond the maximum rate of interest.

Galloway and Shukla's own assumption will become basic for economic theory only after they have explained what importance we are to attach to positive prices when the wage is negative. The failure of their nonreswitching theorem may, however, exempt them now from this task—which is made less easy by the fact that later in the article, they themselves repeatedly dismiss the range of $r$ larger than the maximum, as "outside the pertinent range of the factor price contour."[5]

### III. The Convexity of the Wage Curve

Finally, a remark on a property to which G-S seem to attach considerable importance, namely the convexity to the origin of the "wage curve," (the curve relating the wage and the rate of interest within a given system of production). This convexity implies that in the integrated production of the wage commodity, the value of capital per worker in terms of that commodity falls as the rate of interest rises.[6]

Galloway and Shukla maintain that the constraints ensuring a positive relative price of the commodities at all positive rates of interest also ensure the convexity of the wage curve (see pp. 349; 357–58). This is misleading. Those constraints will not prevent the wage curves from being concave, once their other assumption of a wage advanced at the beginning of the annual production cycle is abandoned in favor of a wage paid post factum.[7]

The convexity of the G-S wage curves is therefore primarily due to the assumption of an "advanced wage" which makes the wage an element of capital, additional to the means of production: as the rate of interest rises this (additional) wage element of capital falls, and thus makes for a fall in the value of capital per worker, i.e., makes for the convexity of the wage curve. The G-S constraints ensuring positive prices play only a subordinate role in this respect: the role of preventing the value of the means-of-production element of capital from rising so much, relative to the wage commodity, as to more than compensate the fall in the wage.[8]

## APPENDIX

### 1. *A Numerical Example*

Consider the following alternative two-commodity systems of production, *I* and *II*:

$$\text{System } I \begin{cases} a_{01} = 1, & \overset{I}{a}_{02} = 1 \\ m_1 = \dfrac{1}{12}, & \overset{I}{m}_2 = \dfrac{1}{6} \\ n_1 = \dfrac{1}{3}, & \overset{I}{n}_2 = \dfrac{1}{6} \end{cases}$$

[5] One such passage reads "However, the second switchpoint lies beyond $r_0$ [the lower of the maximum rates of interest for the two systems concerned] and, therefore, is at a value of $w$ . . . which is negative. Thus it lies outside the pertinent range of the factor price contour" (p. 354; see also pp. 355 and 357).

[6] See the author (1970, p. 419).

[7] This possibility of concave wage curves can be easily seen if we start from the fact that with a wage paid post factum, any indecomposable two-commodity system which gives a convex wage curve when $w$ is in terms of one commodity, say 1, will give a concave curve when $w$ is in terms of the other commodity. Indeed, a convex $w_1(r)$ curve implies a fall in $p_2$, the price of 2 in terms of 1, as $r$ rises: only the fall of $p_2$ can make the given aggregate of the two commodities used as capital in the integrated production of 1 fall in value relative to 1. But, then, a convex $w_1(r)$ implies a rise in $1/p_2$, the price of 1 in terms of 2—and, hence, a rise in the value relative to 2 of the given aggregate of commodities 1 and 2 used as capital for the integrated production of 2; i.e., a convex $w_1(r)$ implies a concave $w_2(r)$ curve. Returning now to G-S constraints on each system of production (see the Appendix) we can note that they are independent of which commodity is used for measuring the wage and prices. What we have seen above then implies that if the constraints are compatible with a convex wage curve, they are by that very fact compatible with a concave curve.

[8] Their constraints on each system of production imply: (a) that each commodity is used in a lower proportion to labor for the production of itself than for the production of the other commodity; and, hence, (b) that the production of one commodity cannot require both means of production in a higher ratio to labor than the production of the other commodity (see fn. 9 below). Both these conditions make for a smaller change in the relative value of the two commodities in the face of any given change in $r$.

$$\text{System } II \begin{cases} a_{01} = 1, & \overset{II}{a_{02}} = \dfrac{92}{91} \\ m_1 = \dfrac{1}{12}, & \overset{II}{m_2} = \dfrac{137}{546} \\ n_1 = \dfrac{1}{3}, & \overset{II}{n_2} = \dfrac{19}{273} \end{cases}$$

where $a_{0i}$, $m_i$, and $n_i$ are, respectively, the inputs of labor, commodity 1, and commodity 2, required to produce a unit of commodity $i$, $(i=1, 2)$. The two systems, which are indecomposable and differ only in the method of production of commodity 2, satisfy Galloway and Shukla's constraints (5) and (6), (p. 349),[9] since:

$$I \begin{cases} a_{01}\overset{I}{m_2} - \overset{I}{a_{02}}m_1 = \dfrac{1}{12} > 0 \\ \overset{I}{a_{02}}n_1 - a_{01}\overset{I}{n_2} = \dfrac{1}{6} > 0 \end{cases}$$

$$II \begin{cases} a_{01}\overset{II}{m_2} - \overset{II}{a_{02}}m_1 = \dfrac{1}{6} > 0 \\ \overset{II}{a_{02}}n_1 - a_{01}\overset{II}{n_2} = \dfrac{73}{273} > 0 \end{cases}$$

From the price equations of each system, we obtain the following relations between $w$ and $r$:

$$\text{(A1)} \quad w^I = \frac{1 - \dfrac{1}{4}(1+r) - \dfrac{1}{24}(1+r)^2}{(1+r) + \dfrac{1}{6}(1+r)^2}$$

$$\text{(A2)} \quad w^{II} = \frac{1 - \dfrac{167}{1092}(1+r) - \dfrac{85}{1092}(1+r)^2}{(1+r) + \dfrac{73}{273}(1+r)^2}$$

Equalizing the right-hand sides of equations (A1) and (A2) we obtain the values $r_1 = 1/3$ and $r_2 = 1/2$ of the rate of interest for which the wage rates in the two systems are the same (i.e., $w_1^I = w_1^{II} = 4/11$; $w_2^I = w_2^{II} = 17/60$). Since for $r=0$, $w^I > w^{II}$, the above result implies that producers will adopt system $I$ for $0 \leq r \leq 1/3$, will then switch to system $II$ for $1/3 \leq r \leq 1/2$, and will finally revert to system $I$ as $r$ increases up to the maximum admissible in the economy $r^I_{max} \cong 1.75$.

## 2. *The Error in Galloway and Shukla's "Proof"*

The coefficients of production of the above numerical example place it in the category which Galloway and Shukla discuss at some length (pp. 355–57) as the third possible under the hypotheses of "Panel B" in their Figure 3. They write the $w$ functions as $w^I = f(r)/\phi(r)$, $w^{II} = f_1(r)/\phi_1(r)$ and discuss the possibility of reswitching in terms of the intersections between the $N = f(r)/f_1(r)$ and $D = \phi(r)/\phi_1(r)$ curves, rather than directly in terms of the intersections between the $w^I$ and $w^{II}$ curves. The decisive error in their argument occurs in the discussion of the curves $\phi(r)$ and $\phi_1(r)$ (the denominators of functions (A1) and (A2), respectively, in the above numerical example), when they fail to notice that for $r=1$, the two curves are tangent to one another (because the two derivatives are $\phi'(-1) = \phi_1'(-1) = a_{01}$) and therefore do not intersect as shown in their Figure 6, p. 356.[10] It follows that the $D = \phi(r)/\phi_1(r)$ curve decreases in the entire interval $r'_{od} < r < -1$ and does not have the minimum shown by G-S in Figures 6 and 7. Accordingly, when $r'_{od} > r'_{on}$, the $D$ curve need

[9] The above constraints on each system can be written as:

$$\frac{m_1}{a_{01}} < \frac{m_2}{a_{02}}; \quad \frac{n_2}{a_{02}} < \frac{n_1}{a_{01}}$$

confirming what was asserted in fn. 8.

[10] Figure 6 is confusing because the points for $r = -1$ of the several functions $\phi(r)$, $\phi_1(r)$, $D$, are all represented as if they occurred in correspondence of point zero of the axis named $r$; moreover, the vertical axis in the figure is incorrectly named $w$ (the same oversights occur in Figures 5 and 7). A more substantial slip occurs in the discussion and representation of curve $N$ as having a single stationary point (p. 355 and Figures 5, 7), whereas the quadratic nature of its derivative should have made clear that if $N$ has one stationary point, then it must have two. Under the hypotheses for which Figures 5 and 7 are drawn, the second stationary point is a maximum occurring for $r > r_{on}$.

not intersect the $N$ curve in the negative range of $r$, and the two curves can intersect twice in the relevant positive interval $0<r<r_{on}$, as happens in my numerical example. (In the G-S notation $r_{on}$, $r'_{on}$ are, respectively, the positive and the negative value of $r$ for which the function $N$ is zero; $r_{od}$, on the other hand, is the (negative) value of $r$ for which $D$ is zero.)

### 3. *The Wage and Prices Cannot be All Positive for $r>r_{max}$*

This proposition can be easily demonstrated for any system with a number $n$ of commodities of which at least one is basic, and each is produced in a separate industry. We can resort to Sraffa's "standard system" where the $n$ industries appear with the unique nonnegative multipliers[11] which ensure that both the net product and the aggregate means of production consist of the same composite "standard commodity." In the standard system the maximum rate of interest compatible with the system of production appears as the ratio between the net product and the aggregate means of production, both measured as quantities of the standard commodity. It follows that for $r>r_{max}$ the amount of profits will consist of a quantity of the standard commodity exceeding the standard net product, and, therefore, the wages in the standard system will be negative together with the wage rate $w_s$ in terms of the standard commodity. If then we refer to the wage $w_i$ in terms of any other commodity $i$, say commodity 1, the equation $w_1=w_s p_s$ shows that $w_1$ will be positive only if $p_s$, the price of the standard commodity in terms of commodity 1, is negative. And since $p_s$ is a weighted average of the prices of the basic commodities in the system, with weights all positive, a negative $p_s$ implies the negativity of at least one of those prices.

In conclusion, for $r>r_{max}$, either the wage will be negative in terms of the chosen value unit or, if positive, at least one of the $(n-1)$ prices will be negative. In either case it will be impossible that both the wage and all prices will be positive.

[11] The multipliers will be positive for the industries producing basic commodities and will be zero for the other industries. (See Sraffa, chs. 4 and 5.)

## REFERENCES

**M. Bruno, E. Burmeister, and E. Sheshinski**, "The Nature and Implications of the Reswitching of Techniques," *Quart. J. Econ.*, Nov. 1966, *80*, 526–53.

**L. Galloway and V. Shukla**, "The Neoclassical Production Function," *Amer. Econ. Rev.*, June 1974, *64*, 348–58.

**P. Garegnani**, "Switching of Techniques," *Quart. J. Econ.*, Nov. 1966, *80*, 554–67.

———, "Heterogeneous Capital, the Production Function and the Theory of Distribution," *Rev. Econ. Stud.*, July 1970, *37*, 407–36.

**D. Levhari**, "A Nonsubstitution Theorem and Switching of Techniques," *Quart. J. Econ.*, Feb. 1965, *79*, 98–105.

**M. Morishima**, "Refutation of the Nonswitching Theorem," *Quart. J. Econ.*, Nov. 1966, *80*, 520–25.

**P. Sraffa**, *Production of Commodities by Means of Commodities*, Cambridge 1960.

# The Neoclassical Production Function: Comment

*By* KAZUO SATO*

The reswitching phenomenon brought to light the fact that the neoclassical production system need not be well behaved. This means that the central proposition in neoclassical production theory, namely the monotonic relation between the round-aboutness of production and the interest rate, need not always hold. There are two ways to cope with this situation. One is to accept what is true, however cumbersome it may be. The other is to look for conditions which rule out "perverse" phenomena in the hope that such conditions are empirically acceptable.

Lowell Gallaway and Vishwa Shukla (abbreviated as G-S henceforth) have taken the second approach and claimed to have discovered a new sufficient condition that rules out the reswitching of techniques in a neoclassical production system, a condition which may be called the capital-intensity condition. Unfortunately, their claim is unfounded; their analysis is faulty. Rather than showing where it has gone astray, I shall present a counterexample that satisfies their capital-intensity condition and yet permits reswitching.[1] Though there are conditions that warrant the neoclassical production system to be well behaved, the capital-intensity condition is not among them.

### I. A Two-Sector Model

Take a well-known two-sector model of a closed economy with the following production structure:

| Sector | | Outputs 1 | Outputs 2 |
|---|---|---|---|
| inputs | labor | $a_{01}$ | $a_{02}$ |
| | 1 | $m_1$ | $m_2$ |
| | 2 | $n_1$ | $n_2$ |

Assume that all inputs are consumed entirely within each unit period and that wages are paid before the period begins. Then, the price equations are given by

$$(1)\quad \begin{cases} p_1 = (1+r)(m_1 p_1 + n_1 p_2 + a_{01} w) \\ p_2 = (1+r)(m_2 p_1 + n_2 p_2 + a_{02} w) \end{cases}$$

where $p$ is prices, $w$ is wage rate, and $r$ is interest rate. Equation (1) is solved to yield equation (2) shown at top of following page, and $w$ may be eliminated by taking the relative commodity price:

$$(3)\quad \frac{p_2}{p_1} = \frac{a_{02} + (a_{01}m_2 - a_{02}m_1)(1+r)}{a_{01} + (a_{02}n_1 - a_{01}n_2)(1+r)}$$

### II. Assertions of Gallaway and Shukla

There are four major assertions, one in each corresponding section of Gallaway and Shukla:

(i) The commodity prices are positive and finite for all positive values of $r$ if and only if:

$$(4)\quad \begin{cases} a_{01}m_2 - a_{02}m_1 > 0 \\ a_{02}n_1 - a_{01}n_2 > 0 \end{cases}$$

(ii) Equation (4) is violated by those counterexamples against nonreswitching in the literature.

(iii) Within the constraints of (4), reswitching cannot occur.

(iv) The factor-price frontier is convex to the origin when (4) prevails.

Equation (4) may be rewritten as

$$(5)\quad n_1/n_2 > a_{01}/a_{02} > m_1/m_2$$

which states that each sector is less capital-intensive in the use of its own product. In this sense, we may call it the capital-intensity condition.[2]

* Professor of economics, State University of New York, Buffalo. I wish to thank Murray Brown for raising an initial suspicion that the main proposition of Gallaway and Shukla is wrong.

[1] A more formal disproof of their "general nonreswitching theorem" is presented in the Appendix.

[2] If $m_1 = m_2 = 0$, the first half of (5) alone is relevant. This truncated capital-intensity condition plays a critical role in this special two-sector growth model as is well known in works of Hirofumi Uzawa and others.

$$(2)\qquad \begin{pmatrix} p_1/w \\ p_2/w \end{pmatrix} = \frac{1}{\left(\frac{1}{1+r} - m_1\right)\left(\frac{1}{1+r} - n_2\right) - m_2 n_1} \begin{bmatrix} \left(\frac{1}{1+r} - n_2\right) a_{01} + n_1 a_{02} \\ \left(\frac{1}{1+r} - m_1\right) a_{02} + n_2 a_{01} \end{bmatrix}$$

### III. The Positivity of Prices

The assertion (i) of G-S is stated in a way that confuses the issue. What they have looked at is the positivity of the relative price $p_2/p_1$. We note two rather obvious points below.

First, as the price system is subject to homogeneity of the first degree, economically relevant prices are $(p_1/w, p_2/w)$ or their reciprocals. What must be positive are these, not their ratio. When they are nonpositive, the technique[3] is not economically viable even though the price ratio is positive. As is well known, with each technique is associated the maximum feasible interest rate $(r^*)$ or the Frobenius root of the technique at which the real wages fall to zero.[4] When $r$ exceeds $r^*$, prices $(p_1, p_2)$ become negative for any positive money wages. They can be made positive only by paying negative wages to workers—an economic impossibility.

Second, when choices are made among alternative techniques, all elements of $(w/p)$ become simultaneously zero for the optimum technique when $r$ reaches the Frobenius root of the entire system. Beyond this point, there is no single technique that keeps all commodity prices positive unless, again, negative wages are paid.

Note that the capital-intensity condition is entirely irrelevant to this analysis and, hence, to the positivity of commodity prices.

### IV. A Counterexample

In their assertion (ii), G-S observed that all counterexamples, so far proposed in the literature, do not satisfy the capital-intensity condition. Suspecting that this failure may be the cause of reswitching, they took great pains to give "a general proof that within the constraints [of (4)] reswitching cannot occur" (p. 351). The proof is wrong. There is nothing better than presenting a counter-example to demolish a faulty proof. I give one such example below.[5]

TABLE 1—TWO TECHNIQUES GIVING RISE TO RESWITCHING

| | I | | | II | | |
|---|---|---|---|---|---|---|
| Sector | 1 | 2 | Ratio | 1 | 2 | Ratio |
| $a_0$ | 1 | 1.5 | .667 | 1 | 1.55 | .645 |
| $m$ | .2 | .4 | .500 | .2 | .5205 | .384 |
| $n$ | .4 | .2 | 2.000 | .4 | .08 | 5.000 |

As everybody (including G-S) has done, take two techniques in which one of the sectors employs the same activity. My example is given in Table 1. Observe that the capital-intensity condition is satisfied in each of the techniques. Taking the first good as the numeraire[6] and writing $y=1/(1+r)$, we obtain the wage-interest equations

$$(6)\quad w/p_1 = \begin{cases} \dfrac{(y-0.6)(y+0.2)}{y+0.4} \\ \text{for technique } I \text{ with } y^*=0.6 \\ \dfrac{y^2-0.28y-0.1922}{y+0.54} \\ \text{for technique } II \text{ with } y^*=0.6002 \end{cases}$$

where $y^*=1/(1+r^*)$. Compute $w/p_1$ for each of the techniques by changing $y$ from 1 to

[3] What I call a technique is referred to as a production system by G-S.

[4] Let $y=1/(1+r)$ and take $g(y)=y^2-(m_1+n_2)y-(m_2n_1-m_1n_2)$, which is the denominator on the right-hand side of (2). By assumption, $g(1)>0$ (i.e., real wages are positive at the zero interest rate). As $g(m_1)<0$ and $g(n_2)<0$, there is one solution $y^*$ to $g(y)=0$ such that $1>y^*>\max(m_1, n_2)$. The corresponding value of $r^*=(1-y^*)/y^*$ is the maximum feasible interest rate.

[5] How such counterexamples can be constructed is shown in the Appendix.

[6] The nonsubstitution theorem ensures that it does not matter which good is chosen as the numeraire.

TABLE 2—WAGE AND INTEREST RATES FOR THE TWO TECHNIQUES

| $y=\frac{1}{1+r}$ | $r$ | $w/p_1$ I | $w/p_1$ II | Dominated by |
|---|---|---|---|---|
| 1.0 | 0 | .3429 | .3427 | Technique *I* |
| .963 | .038 | .3097 | .3097 | Technique *II* |
| .90 | .111 | .2538 | .2540 | |
| .80 | .250 | .1667 | .1670 | |
| .70 | .429 | .0818 | .0821 | |
| .627 | .595 | .0217 | .0217 | |
| .61 | .639 | .0080 | .0079 | Technique *I* |
| .6002 | .666 | .00016 | 0 | |
| .6 | .667 | 0 | — | |

$y^*$. Table 2 shows that the two values are equated at $y=0.963$ and $0.627$ ($r=0.038$ and $0.595$). Technique *I* recurs at the high and low interest rates. Thus, the satisfaction of the capital-intensity condition does not guarantee the nonrecurrence of a technique, definitely disproving the assertion (iii) of G-S.[7]

Table 2 also shows that the two wage-interest curves are almost identical. It is clear that when the production structure is similar across sectors as the capital-intensity condition suggests, techniques tend to have wage-interest curves which are very close to each other. This is the main reason why those who earlier concocted counterexamples all opted for cases that violate the capital-intensity condition. Otherwise, their demonstration would have lost most of its dramatic impact. G-S misread this into their assertion (iii). We may, however, note that when satisfied, the capital-intensity condition definitely reduces the possibility of reswitching.

## V. Further Remarks

In closing,[8] two further remarks may be made. First, are there sufficient conditions that rule out reswitching in the present model? The answer is in the affirmative. For instance, if the capital goods and labor are substitutes in the techniques available, there is no reswitching. (The capital goods themselves may appear as complements with one another.) One of the capital goods has to be complementary with labor to yield the reswitching phenomenon in the model under study. This observation suggests that substitution-complementarity relationships are much more fundamental to the nonperversity of a neoclassical production system than conditions like the capital-intensity condition.

Second, our analysis can be conducted more appropriately when there are more techniques than two. The reswitching phenomenon loses its significance because no single technique is likely to recur (see Michael Bruno et al.). But this does not mean that perverse behavior is absent in the neoclassical system when techniques are continuously available. On the contrary, it is perfectly possible. I have shown elsewhere that sufficient conditions for ruling out perversity in this general case depend also on substitution-complementarity relationships between capital and labor.[9] The capital-intensity condition is not fundamental at all.

[7] This example points to where G-S's mathematics have gone astray. In the form employed by G-S, our two equations (6) can be written as

$$w/p_1 = \begin{cases} \dfrac{f(r)}{\phi(r)} = \dfrac{1-0.4(1+r)-0.12(1+r)^2}{(1+r)+0.4(1+r)^2} \\ \text{for technique } I \\ \dfrac{f_1(r)}{\phi_1(r)} = \dfrac{1-0.28(1+r)-0.1922(1+r)^2}{(1+r)+0.54(1+r)^2} \\ \text{for technique } II \end{cases}$$

An inspection of the four functions given above shows that our example is exactly the case studied in their Panel B of Figure 3 and, in particular, its third possibility that G-S examined at some length. Take $f(r)\phi_1(r) = f_1(r)\phi(r)$. Dividing it through by $(1+r)^2$, we obtain a quadratic equation in $(1+r)$, whose roots are the switching points. G-S assert that one of the roots is outside the range between $r=0$ and $r^*$, which is the root of $f_1(r)=0$, so that there is no reswitching possibility. My example indicates that G-S arrived at this erroneous conclusion through their faulty mathematical reasoning. (G-S, in their Section III, decided that two cases are possible candidates for reswitching: (A) $\phi(r)>\phi_1(r)$ and $f(r)>f_1(r)$ in $0<r<r^*$; (B) $\phi_1(r)>\phi(r)$ and $f_1(r)$ cuts $f(r)$ from above in $0<r<r^*$. This reasoning is mathematically defective. Because techniques *I* and *II* are interchangeable, if (A) is to be inspected, then (A′) $\phi_1(r)>\phi(r)$ and $f_1(r)>f(r)$ cannot be rejected in the way G-S did in a footnote. The same applies to (B).)

[8] The assertion (iv) is examined by others and need not be discussed here.

[9] When all capital goods depreciate at an equal rate, the neoclassical production system can be perverse only if there are some capital goods that are complementary with labor and relatively less labor intensive. This condition is sufficient but not necessary.

## Appendix

### *A General Reswitching Theorem*

This Appendix presents a (somewhat tedious) mathematical demonstration that the capital-intensity condition does not eliminate reswitching. Following G-S, consider "two production systems [techniques] that have in common the technique [activity] for producing commodity one" (p. 351). Let us distinguish them by putting primes to corresponding parameters and variables. We assume that at least one of $n_1$ and $m_2$ is nonzero.[10] Below, let $n_1>0$. We set $w=1$. There are two sets of price equations:

$$\text{(A1)} \quad \begin{aligned} p_1 &= (1+r)(m_1p_1 + n_1p_2 + a_{01}) \\ p_2 &= (1+r)(m_2p_1 + n_2p_2 + a_{02}) \end{aligned}$$

$$\text{(A2)} \quad \begin{aligned} p_1' &= (1+r)(m_1p_1' + n_1p_2' + a_{01}) \\ p_2' &= (1+r)(m_2'p_1' + n_2'p_2' + a_{02}') \end{aligned}$$

Define $\bar{p}_i = p_i' - p_i$, $i=1, 2$. Subtract the corresponding equations of (A1) from (A2) and rearrange terms to obtain

$$\text{(A3)} \quad \bar{p}_2 = \frac{\dfrac{1}{1+r} - m_1}{n_1} \bar{p}_1$$

$$\text{(A4)} \quad \bar{p}_1 = \frac{n_1}{\left(\dfrac{1}{1+r} - m_1\right)\left(\dfrac{1}{1+r} - n_2'\right)} \cdot [(m_2' - m_2)p_1 + (n_2' - n_2)p_2 + (a_{02}' - a_{02})]$$

We observe that reswitching takes place if $\bar{p}_1$ becomes zero more than once between $r=0$ and its maximum feasible value. Note that $\bar{p}_2$ is automatically zero at the same time.

When $\bar{p}_1=0$, we have

$$\text{(A5)} \quad (m_2' - m_2)p_1 + (n_2' - n_2)p_2 + (a_{02}' - a_{02}) = 0$$

Write[11]

$$\text{(A6)} \quad X_1 = -\frac{m_2' - m_2}{a_{02}' - a_{02}} \qquad X_2 = -\frac{n_2' - n_2}{a_{02}' - a_{02}}$$

and $y=1/(1+r)$. Substitute equation (2) in the text into (A5), which is then reduced into a quadratic equation

$$\text{(A7)} \quad h(y) = y^2 - Ay - B = 0$$

where

$$\text{(A8)} \quad \begin{aligned} A &= m_1 + n_2 + a_{01}X_1 + a_{02}X_2 \\ B &= (m_2n_1 - m_1n_2) \\ &\quad + (a_{02}n_1 - a_{01}n_2)X_1 \\ &\quad + (a_{01}m_2 - a_{02}m_1)X_2 \end{aligned}$$

Now assume the capital-intensity condition. If both $X_1$ and $X_2$ are nonnegative, we have $A>0$ and $B>0$, the latter by virtue of that condition. In this case, (A7) has only one positive root. As there is only one switching point at most, no reswitching occurs. Note that both capital goods appear as substitutes for labor between the two techniques.

For reswitching to occur, one of $X_1$ and $X_2$ must be sufficiently negative to make $B<0$. ($X_1$ and $X_2$ cannot be both negative because of (A5).) In other words, one capital good has to be complementary with labor. Reswitching occurs in this case if the two roots of (A7), $y_1$ and $y_2$, both fall between $y^*$ and 1 where $y^*$ ($<1$) is the solution of

$$g(y^*) = y^{*2} - (m_1 + n_2)y^* - (m_2n_1 - m_1n_2) = 0$$

An inspection shows that this is possible if and only if

(a) $h(1) > 0$

(b) $h(y^*) \geqq 0$

(c) $2 > y_1 + y_2 > 2y^*$

(d) $h(A/2) = -A^2/4 - B < 0$

We have

$$\text{(A9)} \quad [(a_{02}n_1 - a_{01}n_2) + a_{01}]X_1 + [(a_{01}m_2 - a_{02}m_1) + a_{02}]X_2 <$$

[10] If both are zero, the two sectors are independent of each other so that there is no reswitching problem.

[11] If $a_{02}' = a_{02}$, we can easily verify that (5) has at most one root by substituting the equation in the text. Hence, no reswitching occurs.

$$1-(m_1+n_2)-(m_2n_1-m_1n_2)\quad(>0)$$

from (a),

$$\text{(A10)}\quad [(a_{02}n_1-a_{01}n_2)+a_{01}y^*]X_1 + [(a_{01}m_2-a_{02}m_1)+a_{02}y^*]X_2 \leqq 0$$

from (b), and

$$\text{(A11)}\quad 2y^*-m_1-n_2 < a_{01}X_1+a_{02}X_2 < 2-m_1-n_2$$

from (c).

These inequalities can be shown graphically in the $(X_1, X_2)$ plane. As

$$\frac{a_{02}n_1-a_{01}n_2}{a_{01}m_2-a_{02}m_1} \gtreqless \frac{(a_{02}n_1-a_{01}n_2)+a_{01}y^*}{(a_{01}m_2-a_{02}m_1)+a_{02}y^*} \gtreqless \frac{(a_{02}n_1-a_{01}n_2)+a_{01}}{(a_{01}m_2-a_{02}m_1)+a_{02}} \gtreqless \frac{a_{01}}{a_{02}}$$

the bounds of (A9), (A10), and (A11) have slopes such that

$$\left.\frac{dX_2}{dX_1}\right|_{(10)} \lesseqgtr \left.\frac{dX_2}{dX_1}\right|_{(9)} \lesseqgtr \left.\frac{dX_2}{dX_1}\right|_{(11)}$$

Equalities hold only when the capital-intensity condition holds with equality sign. In this case, the three bounds are parallel; the intersection of (A10) and (A11) is apparently empty so that there is absolutely no reswitching. Except for this special case, the three bounds differ in slope in such a way that there is always a nonempty area in the plane that satisfies the three inequalities simultaneously. In Figure 1, the upper bound of (A10) and the lower bound of (A11) intersect at point $A$ where the following equality holds (multiply $(1-y^*)$ to (A11), add (A10), and replace inequality by equality signs):

$$[(a_{02}n_1-a_{01}n_2)+a_{01}]X_1 + [(a_{01}n_2-a_{02}m_1)+a_{02}]X_2 = 1-(m_1+n_2)-(m_2n_1-m_1n_2)-(1-y^*)^2$$

This shows that (A9) is satisfied at $A$. Points sufficiently close to $A$ inside the triangle

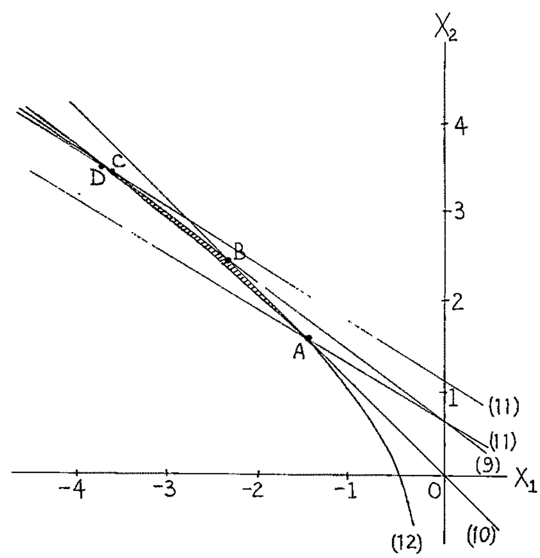



FIGURE 1. TECHNIQUE $I$ OF TABLE 1 RECURS IF TECHNIQUE $II$ IS SUCH THAT $X_1$ AND $X_2$ FALL IN THE AREA $ABCD$
$A(-1.60, 1.60)$, $B(-2.40, 2.40)$, $C(-3.65, 3.50)$, $D(-3.68, 3.52)$.

spanned by the upper bound of (A10) and the lower bound of (A11) satisfy all three inequalities.

For reswitching to occur, however, the roots of (A7) must be real. This is the condition (d) given above, which is

$$\text{(A12)}\quad -\tfrac{1}{4}(a_{01}X_1+a_{02}X_2+m_1+n_2)^2 < (a_{02}n_1-a_{01}n_2)X_1 + (a_{01}m_2-a_{02}m_1)X_2 + (m_2n_1-m_1n_2)$$

As (0, 0) satisfies (A12), the lower bound of (A12) cuts the $X_1$ axis to the left of the origin. We can readily verify that it passes through point $A$ and is tangential to the bound of (A10). As it is convex from above, there is a narrow strip which satisfies (A9), (A10), (A11), and (A12), simultaneously. When $(X_1, X_2)$ falls in this strip, reswitching takes place. Thus, the capital-intensity condition does not eliminate reswitching, even though it narrows down the area in which reswitching is possible.

In the counterexample in the text, we have for technique $I$ the following inequalities which are illustrated in Figure 1:

for (A9): $7X_1 + 8X_2 < 4.8$

for (A10): $X_1 + X_2 \leqq 0$

for (A11): $0.8 < X_1 + 1.5X_2 < 1.6$

for (A12): $-\frac{1}{4}(X_1 + 1.5X_2 + 0.4)^2 < 0.4X_1 + 0.1X_2 + 0.12$

My technique $II$ is obtained for $X_1 = -2.41$, $X_2 = 2.40$, and $a'_{02} - a_{02} = 0.05$. The reader may try other values for himself.

My analysis in this Appendix indicates that reswitching takes place when the two roots of $h(y) = 0$ fall between $y^*$ and 1 where $y^*$ is the largest positive root of $g(y) = 0$. The necessary and sufficient condition for nonreswitching is then the condition that rules it out. One may look for sufficient conditions. What I have accomplished is to show that the capital-intensity condition is not a sufficient condition.

It is easy to demonstrate that substitutability of capital for labor is a sufficient condition for nonreswitching. If there is only one capital good, capital must be a substitute for labor. In this case, either the $m'$ or $n'$ are identically zero. As $B \equiv 0$ in $h(y)$, we have a single positive root for $h(y) = 0$ so that there is no reswitching.

When both goods are used as intermediate inputs, the nonnegativity of $X_1$ and $X_2$ is a sufficient condition. To see this, note that

$$g(1) > 0, \qquad g(m_1) < 0, \qquad g(n_2) < 0,$$
$$1 > y^* > \max(m_1, n_2)$$

We rewrite $h(y)$ as

$$\begin{aligned} h(y) = g(y) &- (a_{01}X_1 + a_{02}X_2)y \\ &- [(a_{02}n_1 - a_{01}n_2)X_1 \\ &+ (a_{01}m_2 - a_{02}m_1)X_2] \end{aligned}$$

so that

$$\begin{aligned} h(m_1) &= g(m_1) - [a_{02}n_1X_1 + a_{01}m_2X_2 \\ &\quad + a_{01}(m_1 - n_2)X_1] \\ h(n_2) &= g(n_2) - [a_{02}n_1X_1 + a_{01}m_2X_2 \\ &\quad + a_{02}(n_2 - m_1)X_2] \end{aligned}$$

Thus, for $(X_1, X_2) \geq (0, 0)$, $h(m_1)$ and/or $h(n_2)$ are negative. Therefore, one root of $h(y) = 0$ is less than $m_1$ and/or $n_2$. This rules out reswitching. In this case, both capital goods appear as substitutes for labor.

When $X_1$ and $X_2$ are opposite in signs, i.e., when one capital good is a substitute for labor and the other is a complement with labor, sufficient conditions for nonreswitching are not readily apparent, but one should be able to derive some of them.[12] For instance, note that

$$\begin{aligned} h(y) = g(y) - y[&\{a_{01} + (a_{02}n_1 - a_{01}n_2)/y\}X_1 \\ &+ \{a_{02} + (a_{01}m_2 - a_{02}m_1)/y\}X_2] \end{aligned}$$

We see that $h(m_1)$ or $h(n_2) < 0$ if the expression in the brackets is nonnegative for $y = m_1$ or $n_2$. A sufficient condition for nonreswitching is that $(X_1, X_2)$ satisfies this inequality.

[12] Joseph Stiglitz, p. 122, introduces one such sufficient condition.

## REFERENCES


M. Bruno, E. Burmeister, and E. Sheshinski, "Nature and Implications of the Reswitching of Techniques," *Quart. J. Econ.*, Nov. 1966, *80*, 526–54.

L. Gallaway and V. Shukla, "The Neoclassical Production Function," *Amer. Econ. Rev.*, June 1974, *64*, 348–58.

K. Sato, "Perversity of the Multisector Neoclassical Production System with no Joint Production," disc. pap. no. 315, State Univ. New York, Buffalo, Sept. 1974.

J. E. Stiglitz, "The Badly Behaved Economy with the Well Behaved Production Function," in J. A. Mirrlees and N. H. Stern, eds., *Models of Economic Growth*, London 1973.

# The Neoclassical Production Function: Reply

*By* LOWELL GALLAWAY AND VISHWA SHUKLA*

The comments of Kazuo Sato and P. Garegnani indicate a technical problem in our original analysis. Its effect is to make it possible for reswitching to occur even though what Sato calls our capital-intensity condition is satisfied, provided that the factor price contours in question are quite similar in nature. Thus, as Sato notes, our condition only "definitely reduces the possibility of reswitching." Unfortunately, all this may be simply a minor exegesis of what is perhaps a trivial question. Let us explain that statement.

Consider the standard case of a full-employment steady-state economy (constant population and labor force) in which alternative two-commodity indecomposable production systems are available. Assume that producers will choose the technique that is "most profitable." Conventionally, this is taken to mean that they will choose the productive technique that will provide the highest profit rate. For example, Luigi Pasinetti states: "Clearly, on grounds of profitability, that technique will be chosen which—for any given wage rate—yields the higher rate of profit" (p. 507). In a micro-economic context one would not quarrel with this proposition. A given investment in capital goods will be more profitable the higher the rate of return on the investment. However, at our highly aggregated (economy wide) level of analysis of the meaning of capital and the nature of the production function, this is not so obvious. At the heart of the Cambridge (England) criticism of neoclassical analysis is the proposition that capital in an aggregate sense can only be measured through the use of some pricing numeraire which itself is functionally related to the level of wage and profit (interest) rates. In this highly interdependent world (illustrated by the two-commodity indecomposable production systems under discussion), it is alleged that the straightforward neoclassical propositions break down because of the interrelationship between the value of capital and the profit (interest) rate. Unfortunately, the analysis upon which this conclusion has been based appears to have ignored the fact that at the aggregate level the phrase "more profitable" has an added dimension that is lacking at the micro-economic level.

This added dimension is the opportunity to employ capital. In the models under analysis the rate of profit and the quantity (in value terms) of capital employed are simultaneously determined. Together, they explain the volume of profits per head associated with a given technique and wage-profit rate combination. Since net output per head for any technique is constant (at the level of the wage rate associated with zero profit rate), profits per head may be calculated by subtracting any wage rate on the factor price contour from the zero profit wage. We have done this for Sato's example (Garegnani's would do just as well) and the results are shown in Table 1. They are fascinating in that they indicate that *at any profit rate the most profitable technique from the standpoint of profits per head is technique I*. But, how important is this? Quite; in a macro-model that assumes 1) full employment and 2) that capital is used solely for the purpose of producing the commodities included in the model.[1] Together, these conditions imply that profits per head are a measure of the total profits associated with the use of any technique. This means that switching from technique *I* to technique *II* in Sato's example as the profit rate falls below .595 involves adopting a production system that yields

* Ohio University.

[1] Actually, in the systems under consideration capital goods are produced simultaneously with consumption goods. In a steady-state equilibrium their current production is just sufficient to permit the replacement of capital goods that have depreciated.

TABLE 1—WAGE RATE-PROFIT RATE COMBINATIONS AND TOTAL PROFITS, TWO ALTERNATIVE TWO-COMMODITY INDECOMPOSABLE PRODUCTION SYSTEMS

| Profit Rate ($r$) | Wage Rate ($w$) | | Total Profits ($rk$) | |
|---|---|---|---|---|
| | Technique *I* | Technique *II* | Technique *I* | Technique *II* |
| .000 | .3429 | .3427 | .0000 | .0000 |
| .038[a] | .3097 | .3097 | .0332 | .0330 |
| .111 | .2538 | .2540 | .0891 | .0887 |
| .250 | .1667 | .1670 | .1762 | .1757 |
| .429 | .0818 | .0821 | .2611 | .2606 |
| .595[a] | .0217 | .0217 | .3212 | .3210 |
| .639 | .0080 | .0079 | .3349 | .3348 |
| .666 | .00016 | .0000 | .34274 | .3427 |
| .667 | .0000 | — | .3429 | — |

[a] Conventionally defined switchpoint.

less real income to producers, despite the higher profit rates associated with technique *II*. The reason is simple and has been noted previously in the literature:[2] the value of the capital employed with technique *II* is more than enough less than that used in technique *I* to offset the higher profit rates that go with *II*.

What really is involved in moving from technique *I* to technique *II* as $r$ falls below .595? A comparison of the steady-state equilibrium positions associated with the two techniques reveals that the transition amounts to adopting a system of production that produces and employs less capital per head (in value terms) at a higher profit rate with less output per head and a lower level of real income for producers. The puzzling thing is why producers would switch techniques under these conditions. The forces of competition? We think not. At a given wage rate the option is always open to use technique *I* with its lower profit rates but higher level of output per head and producers' real income. No, it would seem that if by the forces of competition we mean the result of maximizing behavior by large numbers of producers, *in the absence of what we prefer to call a "profit rate illusion," technique II would never be adopted and reswitching would not occur in the context of these models.* What we are saying is that the best interests of producers (in terms of economic welfare) are not synonymous with the simple maximization of profit rates when dealing with this level of aggregation. Single-minded dedication to the maximization of the rate of profit can lead producers to adopting courses of behavior that so limit the opportunities for employing capital (with the fixed labor force implicit in the models) that lower real incomes will be their reward.

There is a further implication of this argument, namely, that if the problem is one of choosing among already known alternative two-commodity indecomposable production systems, not only will there be no reswitching, but, in addition, *there will be no switching of techniques.* One technique alone will dominate all others and that technique will be the one that has the greatest output per head. Whatever wage rate exists, profits are maximized by adopting that technique. What does this imply about the aggregate production function in a world of economy wide two-commodity indecomposable production systems? Simply this; it is the system that will yield the highest output per head. As a general rule, that system will generate a non-linear factor price contour which, given our capital-intensity condition, will be well ordered in the sense that higher wage rates and lower profit rates will be associated with

[2] For example, Pasinetti, p. 514, notes this phenomenon and calls it a "remarkable" result because it produces results that conflict with the neoclassical version of things. Of course, he continues to assume that the appropriate object of maximization is the profit rate.

higher capital-labor ratios where capital is measured in value terms.[3]

[3] Another way of putting this is to note that any point on the conventionally defined factor price frontier that is not associated with the technique that yields maximum output per head is non-Pareto optimal in that it is possible to move to another technique and improve the welfare of some members of the society without damaging the welfare of anyone else. For example, profit income may be increased without reducing wage income or, alternatively, workers could hire producers (paying them slightly more than the income they are earning using an inferior technique) to use the dominant technique instead and pocket the gains for themselves.

## REFERENCES

**L. Gallaway and V. Shukla**, "The Neoclassical Production Function," *Amer. Econ. Rev.*, June 1974, *64*, 348–58.

**P. Garegnani**, "The Neoclassical Production Function: Comment," *Amer. Econ. Rev.*, June 1976, *66*, 424–27.

**L. L. Pasinetti**, "Changes in the Rate of Profit and Switches of Techniques," *Quart. J. Econ.*, Nov. 1966, *80*, 503–17.

**K. Sato**, "The Neoclassical Production Function: Comment," *Amer. Econ. Rev.*, June 1976, *66*, 428–33.

# A Note on the U Hypothesis Relating Income Inequality and Economic Development

By SHERMAN ROBINSON*

A common empirical finding in the analysis of countries which have undergone economic development is that income distribution first became more unequal, and only in the later phase did it become more equal. This empirical observation has also been seen in modern developing countries—at least the increasing inequality phase—and has acquired the force of economic law. It has a name: the U hypothesis. A number of different economic explanations for the relationship have been presented involving a variety of factors such as productivity changes, differential savings behavior, exploitation of workers, and so forth.[1]

The purpose of this note is to demonstrate that the U hypothesis can be derived from a very simple model with a minimum of economic assumptions. One need only assume that the economy can be divided into two sectors with different sectoral income distributions and that there is a monotonic increase in the relative population of one of the sectors over time. These assumptions seem empirically unexceptional for a country undergoing economic development and are consistent with many models of development such as the Lewis-Fei-Ranis surplus labor models, dual economy models, and the more recent Harris-Todaro migration models.[2] The U hypothesis will be seen to be a necessary implication of the workings of such models.

Assume that the economy is divided into two sectors with different income distributions. The *log* mean and *log* variance of income in the two sectors are given by $Y_1$ and $Y_2$ and $\sigma_1^2$ and $\sigma_2^2$, respectively. Define the population shares of the two sectors as $W_1$ and $W_2$ with:

$$W_1 + W_2 = 1 \tag{1}$$

The overall *log* mean income is given by:

$$Y = W_1Y_1 + W_2Y_2 \tag{2}$$

and the overall *log* variance is given by:

$$\sigma^2 = W_1\sigma_1^2 + W_2\sigma_2^2 + W_1(Y_1 - Y)^2 + W_2(Y_2 - Y)^2 \tag{3}$$

The *log* variance is itself an increasing measure of income inequality. One need not use *log* means and *log* variances, but they are convenient since the *log* variance is a commonly used inequality measure. The arithmetic mean and variance would also do—the algebra is exactly the same.

Assuming that the within-sector distributions remain unchanged over time ($\sigma_1^2$, $\sigma_2^2$, $Y_1$, and $Y_2$ are constant), then from equation (3) inequality is a function only of sectoral population shares and overall *log* mean income. By equation (2), the overall *log* mean is itself a function of sectoral population shares.

Assume that sector 1 is the sector whose relative population share is increasing. Then, substituting (1) and (2) into (3) and doing a bit of algebra, one finally gets:

$$\sigma^2 = AW_1^2 + BW_1 + C \tag{4}$$

where
$$A = -(Y_1 - Y_2)^2$$
$$B = (\sigma_1^2 - \sigma_2^2) + (Y_1 - Y_2)^2$$
$$C = \sigma_2^2$$

If one assumes that the *log* mean incomes are different in the two sectors, then in-

* Assistant professor of economics, Princeton University. I am grateful to Irma Adelman for comments on an earlier draft and to Charles Frank and Richard Webb for useful discussions. They, of course, bear no responsibility for any remaining errors.

[1] See Irma Adelman and Cynthia Morris, p. 188, Simon Kuznets (1955, 1966), and Felix Paukert.

[2] See Kuznets (1966, 1971), W. Arthur Lewis, John Fei and Gustav Ranis, and John Harris and Michael Todaro.

equality is a quadratic function of $W_1$. Since $A<0$, the parabola has a maximum. As $W_1$ increases, inequality first increases, reaches a maximum, then decreases—precisely the U hypothesis.

There is one possible problem. Since by assumption $0 \leq W_1 \leq 1$, it is possible that the maximum value of $\sigma^2$ occurs for a value of $W_1$ outside the zero to one range. Setting the first derivative of (4) equal to zero, the maximum value of $\sigma^2$ occurs when $W_1$ is equal to $\hat{W}_1$:

$$(5) \qquad \hat{W}_1 = \frac{\sigma_1^2 - \sigma_2^2}{2 \cdot (Y_1 - Y_2)^2} + \frac{1}{2}$$

Thus, the more equal are the *log* variances, and the more different are the *log* mean incomes, the closer is $\hat{W}_1$ to 1/2. Empirically, one would expect a much greater difference in *log* mean incomes than in *log* variances and so expect $\hat{W}_1$ to be in the range zero to one.

Some properties of these equations are interesting. The U hypothesis in no way depends on which sector has the higher income. If total income is to rise then $Y_1$ must be greater than $Y_2$, but the U hypothesis depends only on their being different. It also does not matter which sector has the more unequal distribution of within sector income. The difference between $\sigma_1^2$ and $\sigma_2^2$ affects $\hat{W}_1$, but not the existence of the U. If $\sigma_1^2<\sigma_2^2$, then it will take longer for the distribution to start becoming more equal (for a given rate of change of sector population shares), but the turning point exists. It is interesting that even if people are moved from a sector with relatively more equality to one with less, the overall distribution will still become more equal.

A simple example will given an idea of the orders of magnitude involved. Assume the following parameter values:[3]

$$Y_1 = 6.91 \qquad Y_2 = 4.61$$
$$\sigma_1^2 = .40 \qquad \sigma_2^2 = .20$$

[3] The within-group *log* variances are based roughly on data from studies of South Korea and Turkey. The group mean incomes corresponding to the *log* means are rough guesses and the effects of varying them are explored below.

TABLE 1—DISTRIBUTION STATISTICS: VARIOUS GROUP POPULATION SHARES

| $W_1$ | $Y$ | Geometric Means | $\sigma^2$ | Gini Coefficient |
|---|---|---|---|---|
| .10 | 4.84 | 126. | .70 | .44 |
| .20 | 5.07 | 159. | 1.09 | .54 |
| .40 | 5.53 | 252. | 1.55 | .62 |
| .52 | 5.80 | 331. | 1.63 | .63 |
| .60 | 5.99 | 399. | 1.59 | .63 |
| .80 | 6.45 | 633. | 1.21 | .56 |
| .90 | 6.68 | 796. | .86 | .49 |

*Note:* The geometric mean is exp ($Y$). The Gini coefficient is calculated from the *log* variance assuming a lognormal distribution. The overall *log* variance $\sigma^2$ is calculated from equation (4). $(\sigma_1^2-\sigma_2^2)=.20$.

The values of $Y_1$ and $Y_2$ correspond to geometric mean incomes of 1000 and 100, respectively. If one assumes a lognormal distribution, then the arithmetic mean income is a function of the *log* mean and *log* variances and the Gini coefficient is a simple function of *log* variance.[4] The corresponding arithmetic mean incomes are 1221 and 111 and the within-sector Gini coefficients are 0.35 and 0.25 for distributions one and two, respectively. Statistics for the overall distribution given different values of $W_1$ are presented in Table 1. The maximum value of $\sigma^2$ occurs when $\hat{W}_1=.52$. The U relationship between inequality and development appears very distinctly in the example. It is interesting that even when $W_1=.10$ or $W_1=.90$, $\sigma^2$ is significantly different from $\sigma_2^2$ or $\sigma_1^2$.

During the course of economic development, one also might expect the mean income differences of various groups to narrow, perhaps after an initial rise. Table 2 indicates how inequality varies as the ratio of group mean incomes change. If one assumes that group population shares do not change

[4] For a lognormal distribution with *log* variance $\sigma^2$, the Gini coefficient is given by $G=2[\int_{-\infty}^{x} N(0,1)]-1.0$ where $x=\sigma/\sqrt{2}$ and $N(0, 1)$ is the standard normal distribution. The arithmetic mean is equal to exp $(Y+0.5\,\sigma^2)$ where $Y$ is the *log* mean. See J. Aitchison and J. A. C. Brown. Note that the use of this formula for the overall distribution assuming that the within-group distributions are lognormal is only an approximation since the sum of two lognormal distributions is not necessarily lognormal.

TABLE 2—DISTRIBUTION STATISTICS: VARIOUS GROUP INCOME RATIOS

| Income Ratio | $Y_1-Y_2$ | $\hat{W}_1$ | $\sigma^2$ | Gini Coefficient |
|---|---|---|---|---|
| 10 | 2.30 | .52 | 1.63 | .63 |
| 9 | 2.20 | .52 | 1.51 | .62 |
| 8 | 2.08 | .52 | 1.38 | .59 |
| 7 | 1.95 | .53 | 1.25 | .57 |
| 6 | 1.79 | .53 | 1.11 | .54 |
| 5 | 1.61 | .54 | 0.95 | .51 |
| 4 | 1.39 | .55 | 0.79 | .47 |
| 3 | 1.10 | .58 | 0.61 | .42 |
| 2 | 0.69 | .71 | 0.44 | .36 |

*Note:* Income ratio is the ratio of geometric mean incomes, exp $(Y_1)$/exp $(Y_2)$. The variance $\sigma^2$ is calculated from equation (4) for the corresponding turning point value of $W_1$. $(\sigma_1^2-\sigma_2^2)=.20$.

but only that the spread of group mean incomes first widens and then later narrows, one also gets a significant U relationship.

It is possible to compare the relative magnitudes of the effect on inequality of changes in group mean incomes and of changes in group population shares. Define $R=Y_1-Y_2$. The logarithm of the ratio of group mean incomes is $R$. If $Y_1>Y_2$, then $R>0$. Equation (4) can be rewritten as:

$$(6) \quad \sigma^2 = -R^2W_1^2 + [(\sigma_1^2-\sigma_2^2)+R^2]W_1 + \sigma_2^2$$

Take the total derivative of (6) with respect to changes in $R$ and $W_1$:

$$(7) \quad d\sigma^2 = D\,dR + E\,dW_1$$

where $D = 2RW_1W_2$

$$E = R^2(1-2W_1) + (\sigma_1^2-\sigma_2^2)$$

At the turning point $\hat{W}_1$ in equation (5), $E=0$ and only changes in relative mean incomes affect inequality. If $W_1=1/2$, $E=(\sigma_1^2-\sigma_2^2)$ and does not depend on $R$. For a given $R$, $E$ is relatively small when $W_1$ is in the neighborhood of 1/2. Also, given $R$, $D$ is largest when $W_1=W_2=1/2$.

Table 3 gives some values of $D$ and $E$ for various values of $W_1$ and $R$. A change in $R$ of 0.1 corresponds roughly to a decrease of 1.0 in the ratio of mean incomes (say, from

TABLE 3—COEFFICIENTS OF THE TOTAL DERIVATIVE $d\sigma^2=DdR+EdW_1$

| $W_1$ | $R$ | $0.1D$ | $0.01E$ |
|---|---|---|---|
| .10 | 2.30 | .041 | .044 |
| .10 | 1.61 | .029 | .023 |
| .10 | 0.69 | .012 | .006 |
| .20 | 2.30 | .074 | .034 |
| .20 | 1.61 | .052 | .018 |
| .20 | 0.69 | .022 | .005 |
| .40 | 2.30 | .111 | .013 |
| .40 | 1.61 | .077 | .007 |
| .40 | 0.69 | .033 | .003 |

*Note:* $R$ corresponds to income ratios of 10 ($R=2.30$), 5 ($R=1.61$), and 2 ($R=.69$). $(\sigma_1^2-\sigma_2^2)=.20$.

10 to 9). A change in $W_1$ of 0.01 (one percentage point) seems roughly comparable to a change in $R$ of 0.1. "Comparable" means that a country undergoing rapid structural change might achieve such a change in a year or two, although clearly a change in the income ratio from 10 to 9 is far easier than a change from 3 to 2.[5]

It is clear from both Tables 1 and 3 that given $R$, the U relation is very flat around $W_1=1/2$. Changes in $W_1$ from 0.40 to 0.60 have almost no effect on inequality. At the ends where $W_1=0.2$ or 0.1, the effect of changes in $W_1$ are comparable to changes in $R$. In the development process as it has actually occurred, $R$ has typically first increased and then later decreased. The two effects have thus probably reinforced one another in generating the U relationship. If $R$ stays relatively constant during the "middle" period of development, one would find that a country might spend a long time in the trough of the U.

The implication of this exercise is that if the two-sector models that many development economists use are valid, then one should expect that a developing country, in the absence of explicit countervailing policies, will have increasing or unchanged income inequality for a relatively long period. The result is also consistent with the em-

[5] For evidence on the rate of structural change in less developed countries, see Kuznets (1966, 1971) and the author.

pirical findings of Adelman and Morris. Of course, the model is very simple and ignores any systematic changes in the within-group distributions over time. It does, however, trace out the effects of simple structural shifts on income inequality, effects which are significant and should not be neglected in more complicated and more theoretically satisfying models.

## REFERENCES


I. Adelman and C. T. Morris, *Economic Growth and Social Equity in Developing Countries,* Stanford 1973.

J. Aitchison and J. A. C. Brown, *The Lognormal Distribution,* Cambridge 1957.

J. C. H. Fei and G. Ranis, *Development of the Labor Surplus Economy: Theory and Policy,* Homewood 1964.

J. R. Harris and M. Todaro, "Migration, Unemployment, and Development: A Two Sector Analysis," *Amer. Econ. Rev.,* Mar. 1970, *60,* 126–43.

S. Kuznets, *Modern Economic Growth: Rate, Structure and Spread,* New Haven 1966.

———, *Economic Growth of Nations,* Cambridge, Mass. 1971.

———, "Economic Growth and Income Inequality," *Amer. Econ. Rev.,* Mar. 1955, *45,* 1–28.

W. A. Lewis, "Economic Development with Unlimited Supplies of Labour," *Manchester Sch. Econ. Soc. Stud.,* May 1954, *22,* 139–91.

F. Paukert, "Income Distribution at Different Levels of Development," *Int. Lab. Rev.,* Aug.-Sept. 1973, *108,* 97–125.

S. Robinson, "Sources of Growth in Less Developed Countries: A Cross-Section Study," *Quart. J. Econ.,* Aug. 1971, *85,* 391–408.

# Intermediate Products and the Theory of International Trade: A Generalization of the Pure Intermediate Good Case

*By* JAMES RIEDEL*

Empirical studies of neoclassical, factor-proportions trade theory have long recognized that hypotheses concerning relative factor intensity must be defined in terms of gross factor intensities (accounting for intensities of intermediate products) rather than simply in net factor-intensity terms (i.e., at the last stage of production only).[1] Nevertheless, development of the pure theory of international trade has generally neglected intermediate products. The justification for this convenience has in part been based on a study by Jaroslav Vanek which demonstrated that the introduction of interindustry flows in a two-good model did not affect the traditional theorems (Rybczynski, Stolper-Samuelson, or Heckscher-Ohlin) defined in net factor-intensity terms. However, in a recent study, Raveendra Batra and Francisco Casas demonstrated that the introduction of a "pure" intermediate commodity—one that is produced solely for use in the production of final goods—complicates the issue considerably. They show, in fact, that " . . . most of the traditional trade theorems [defined in net terms] may not hold without additional provisions if such intermediate products are introduced in the model . . . " (p. 297). The basic difference between the Batra-Casas model with pure intermediate products and Vanek's interindustry flow model is the possibility of a conflict between the net and gross factor-intensity ranking of final products.[2]

The conditions which Batra and Casas demonstrate are necessary in order to ensure a uniform ranking of final commodities according to net and gross factor intensity were derived under the explicit assumption that the intermediate good is totally supplied from domestic production. The purpose of the present paper is to generalize the Batra-Casas model to allow the intermediate good to be imported in some proportion.[3] The analytical approach of the present paper is quite different from that taken by Batra and Casas, however. Here the problem is treated as a problem of *ex post* factor-intensity measurement, although the underlying model and necessary conditions to ensure its proper functioning are implicitly the same as described by Batra and Casas. In Section I a measure of gross factor intensity appropriate in an economy importing intermediate inputs is developed. In the following section this measure is applied to the Batra-Casas model, under alternative assumptions regarding the supply source of the intermediate good. It is shown that if some proportion of the intermediate good is imported, then net and gross factor-intensity rankings can be assured to be uniform if and only if—in addition to the

* Research associate, Institut für Weltwirtschaft an der Universität Kiel. This paper reports research undertaken with the "Sonderforschungsbereich Nr. 86, Weltwirtschaft und internationale Wirtschaftsbeziehungen" with financial support from the Deutsche Forschungsgemeinschaft. I acknowledge, without implicating in remaining errors, the helpful comments of R. Banerji, F. R. Casas, and L. L. Wegge.

[1] See Wassily Leontief (1953, 1956).

[2] Analyzing a model different from Batra and Casas only with respect to the assumed structure of production, Alok Ray arrived, not surprisingly, at the very same conclusion.

[3] In a recent paper, Albert Schweinberger examines the possibility of a conflict in net and gross factor-intensity ranking in a model containing a traded pure intermediate product. Schweinberger, for reasons not stated, chose to ignore Batra and Casas' proof of absence of the indeterminancy problem (three commodities, two primary inputs), avoiding the problem instead by assuming that trade does not lead to factor price equalization. This assumption allows that although each country produces not more than two of the three goods in equilibrium, the world as a whole will produce all three. The relaxation of the nonspecialization assumption constitutes a quite different model from that analyzed by Batra and Casas, however.

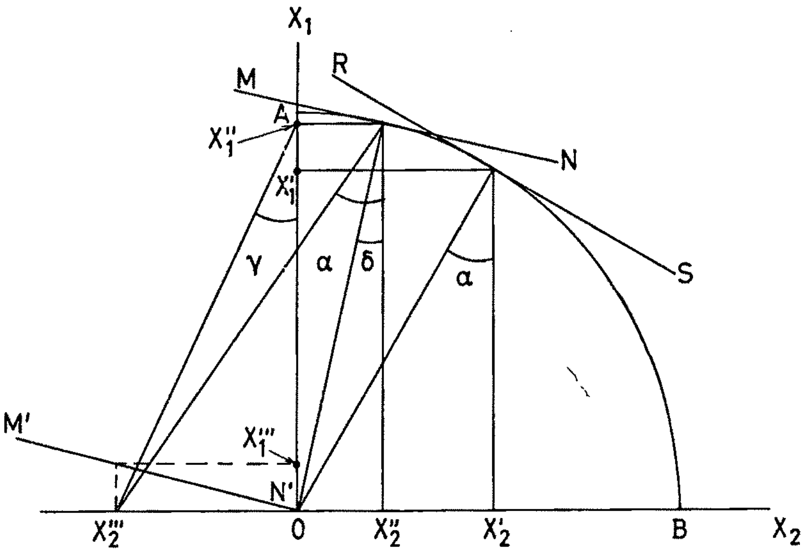



FIGURE 1

Batra-Casas conditions—the proportion of the intermediate good imported for use in producing the two final goods is the same for both goods. In addition, it is shown that if the pure intermediate good is totally imported no complications arise, the Batra-Casas model becomes essentially the same as Vanek's model.

## I. Factor-Intensity Measurement with Imported Intermediate Inputs[4]

Clearly there is no a priori reason to assume any given mix between imported and domestically supplied intermediate goods. The optimal mix will depend on relative prices. Consider the two-good case for simplicity (see Figure 1). Curve $AB$ describes the production possibility frontier between $X_1$ and $X_2$. Assume the production of $OX_1'$ of $X_1$ requires $OX_2'$ of $X_2$ as an intermediate input. The technical coefficient, $a_{21}(X_{21}/X_1)$, is thus equal to the *tan* $\alpha$. For any set of relative prices $(P_2/P_1)$ less than $RS$, such as $MN$ for example, the optimal solution will require some proportion of $X_2$ to be imported.

Given the set of relative prices described by $MN$, the optimal solution will demand that the production of $X_1$ be increased from $X_1'$ to $X_1''$. The production of $OX_1''$, however, requires $X_2'''X_2''$ of $X_2$, of which only $OX_2''$ can be supplied domestically, the remainder $0X_2'''$ necessarily being imported in exchange for $X_1(0X_1'''$ of $X_1)$.[5] Under this set of relative prices, the domestic input coefficient $(d_{ij})$ is expressed by *tan* $\delta$, and the imported input coefficient $(m_{ij})$ by the *tan* $\gamma$.

The importation of intermediate inputs has important implications for the measurement of gross factor intensity.[6] The orthodox measure of gross factor intensity, following Leontief, is computed by taking a weighted average of per unit factor requirements at every stage in production, the weights being the direct and indirect per unit input requirements at each stage of production:

$$k_j^* = \frac{\sum_i C_{Ki} r_{ij}}{\sum_i C_{Li} r_{ij}} \tag{1}$$

[4] The gross factor-intensity measure described in this section was presented in my 1975 paper and applied to the Taiwan economy in an attempt to analyze the rationale of a "footloose" industrial structure.

[5] $M'N'$, parallel to $MN$, defines the terms of trade between $X_1$ and $X_2$.

[6] Note, gross factor intensity is measured in gross output terms (labor and capital per unit output) as is the usual convention.

where $C_{Ki}$ and $C_{Li}$ are the direct capital and labor output ratios, respectively, and where $r_{ij}$ are elements of the inverted Leontief technology matrix $[I-A]^{-1}$. It should be obvious that the standard measure is inappropriate in an economy which imports intermediate inputs, since the $[I-A]^{-1}$ matrix (which weights direct factor intensities at each stage of production) explicitly assumes that all intermediates are domestically produced. If some proportion of the intermediate inputs is imported, however, what is relevant is not the factor requirements of producing the input directly, but rather the factor requirements of producing the goods which are exported in exchange for the imported inputs. Gross factor intensity in an open economy importing intermediate goods should reflect the factor requirements at the last stage of production, in producing the intermediate inputs supplied domestically and in producing the exports which are exchanged for intermediate inputs imported.

If inputs are imported, the domestically produced inputs required directly and indirectly per unit final demand are given by the inverted Leontief matrix of domestic transactions[7]

$$[I - A - M]^{-1} = [I - D]^{-1} = \{s_{ij}\}$$

where $M$ is a matrix of per unit imported inputs $(m_{ij})$.

The labor and capital required per unit output at the last stage of production and in producing the domestic intermediate inputs for the production of commodity $j$ are

$$C^*_{Li} = \sum C_{Li}s_{ij} \tag{2a}$$

$$C^*_{Ki} = \sum_i C_{Ki}s_{ij} \tag{2b}$$

Interindustry transactions in imported inputs are omitted because the demand for imported inputs does not directly affect employment or capital investment in the domestic economy. Nevertheless, imported inputs are not freely available to the economy, much less the domestic producer; they can be acquired (in equilibrium) only through the exchange of domestically produced goods and services, which in turn involve a resource cost. Thus, gross factor intensity in an economy which imports intermediate inputs is accurately measured by the sum of 1) the labor and capital employed in producing domestic inputs and at the final stage of production ($C^*_{Li}$ and $C^*_{Ki}$), and 2) the labor and capital cost implicit in earning the foreign exchange (exporting) with which to purchase imported inputs required directly and indirectly in the production process.

The labor and capital cost of earning one unit of foreign exchange (in equilibrium) is the labor and capital required to produce one unit of exports, which in turn can be defined as the average labor and capital requirements per unit output in each sector of the economy weighted by the distribution of exports from each sector. The labor and capital required at the last stage of production and in the domestic production of inputs which go into exports is thus $C^*_{LF}$ and $C^*_{KF}$, respectively, as defined.

$$C^*_{LF} = \sum_j \left[ \sum_i C_{Li}s_{ij} \right] e_j$$

$$C^*_{KF} = \sum_j \left[ \sum_i C_{Ki}s_{ij} \right] e_j$$

The coefficient $e_j$ is the $j$th sector's share in total exports. Of course the production of exports itself requires imported inputs. If $M^*_j$, defined as[8]

$$M^*_j = \sum_i M_i s_{ij}$$

is the direct and indirect import requirement per unit output of commodity $j$, then

$$M^*_F = \sum_j \left[ \sum_i M_i \cdot s_{ij} \right] e_j = \sum_j M^*_j \cdot e_j$$

is the direct and indirect import requirement per unit export. Thus to produce one unit of exports we need $C^*_{LF}$ units of labor and $C^*_{KF}$

[7] According to standard notation:

$$[D] = \{d_{ij}\}$$
$$d_{ij} = a_{ij} - m_{ij}$$

[8] Note $M_j (= \sum_i m_{ij})$ is the total direct import requirement in the production of the $j$th good.

units of capital at the last stage of production and for domestically produced inputs; and we need $C^*_{LF}M^*_F$ and $C^*_{KF}M^*_F$ of labor and capital, respectively, to produce additional exports to finance the imports which were employed in the original production of one unit of exports. In addition, we recognize that the additional exports (required to finance the imports used in the first round) also require imported inputs. In the first round $M^*_F$ units of imports (foreign exchange = exports) are required; thus in the second round $M^*_F \cdot M^*_F$ units of imported inputs are required, which in turn will entail the employment of $C^*_{LF}(M^{*2}_F)$ labor and $C^*_{KF}(M^{*2}_F)$ capital in the production of exports with which to finance these additional imports. The second round of additional exports, likewise, requires imported inputs $(M^{*2}_F \cdot M^*_F)$ and consequently more exports and hence the employment of still more labor and capital, and so on. The sum of all labor and capital required in the production of one unit of exports (i.e., foreign exchange) is thus

$$\text{(3a)}\quad C^*_{LF} + C^*_{LF}M^*_F + C^*_{LF}(M^{*2}_F) + C^*_{LF}(M^{*3}_F) + \ldots + C^*_{LF}(M^{*n}_F) = \frac{C^*_{LF}}{1 - M^*_F} \qquad n \longrightarrow \infty$$

$$\text{(3b)}\quad C^*_{KF} + C^*_{KF}M^*_F + C^*_{KF}(M^{*2}_F) + C^*_{KF}(M^{*3}_F) + \ldots + C^*_{KF}(M^{*n}_F) = \frac{C^*_{KF}}{1 - M^*_F} \qquad n \longrightarrow \infty$$

since $0 \leq M^*_F \leq 1$.

For any given commodity $j$, therefore, the factor intensity of production as measured by total factor requirement per unit output is given by

$$\text{(4)}\quad k^{**}_j = \frac{\sum_i C_{Ki}s_{ij} + M^*_j\left[\dfrac{C^*_{KF}}{1 - M^*_F}\right]}{\sum_i C_{Li}s_{ij} + M^*_j\left[\dfrac{C^*_{LF}}{1 - M^*_F}\right]}$$

which expresses the two components of total factor cost in an open economy: 1) labor and capital cost at the last stage of production and producing domestic inputs; and 2) labor and capital cost implicit in earning the foreign exchange (fraction of one unit of foreign exchange) with which to purchase imported inputs required directly and indirectly in the production of commodity $j$.

## II. Factor-Intensity Ranking Under Alternative Assumptions Regarding the Supply Source of Intermediate Products

### A. *The Intermediate Product Totally Supplied from Domestic Production*

The assumption that the pure intermediate is totally supplied from domestic production was explicit in the Batra-Casas model, which consisted of three commodities, two final products ($X_1$ and $X_2$), and one intermediate product ($X_3$), and assumed a production structure given by the Leontief matrix as

$$[A] = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ a_{31} & a_{32} & 0 \end{bmatrix}$$

It is clear that if the intermediate inputs are totally supplied domestically the gross factor-intensity measure in equation (4) becomes the same as the orthodox measure in equation (1).[9] Applying this measure to the Batra-Casas model yields

$$k^*_1 = \frac{C_{K1} + C_{K3}a_{31}}{C_{L1} + C_{L3}a_{31}} = \frac{k_1C_{L1} + k_3C_{L3}a_{31}}{C_{L1} + C_{L3}a_{31}}$$

$$k^*_2 = \frac{C_{K2} + C_{K3}a_{32}}{C_{L2} + C_{L3}a_{32}} = \frac{k_2C_{L2} + k_3C_{L3}a_{32}}{C_{L2} + C_{L3}a_{32}}$$

$$k^*_3 = \frac{C_{K3}}{C_{L3}}$$

where $k_i$ is the *net* factor intensity $(K/L)$ of product $i$. The central question upon which the Batra-Casas analysis hinges is,

[9] Note:

$$[I - A]^{-1} = \{r_{ij}\} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ r_{31} & r_{32} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ a_{31} & a_{32} & 1 \end{bmatrix}$$

"... if the first commodity is *apparently* capital intensive relative to the second commodity (that is, if $k_1 > k_2$), then will it be *truly* capital intensive (that is, $k_1^* > k_2^*$) ... ?" (p. 304, emphasis added). The condition is that

$$\frac{k_1 C_{L1} + k_3 C_{L3} a_{31}}{C_{L1} + C_{L3} a_{31}} > \frac{k_2 C_{L2} + k_3 C_{L3} a_{32}}{C_{L2} + C_{L3} a_{32}}$$

or

$$C_{L1}C_{L2}(k_1 - k_2) + a_{31}C_{L2}C_{L3}(k_3 - k_2) + a_{32}C_{L1}C_{L3}(k_1 - k_3) > 0$$

which Batra and Casas showed, holds only if

(i) $k_1 > k_3 > k_2$

or

(ii) $k_1 > k_2 > k_3$ and $a_{32}/C_{L2} \geq a_{31}/C_{L1}$

### B. *The Intermediate Product Totally Imported*

The Batra-Casas model is now examined under the assumption that the intermediate is totally imported. As a matter of convenience it is assumed that only $X_1$ is exported, though it makes no difference if both final goods are exported in exchange for the intermediate. Under the assumption that the intermediate is totally imported, the inverted domestic Leontief matrix is

$$[I - D]^{-1} = \{s_{ij}\} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The per unit direct and indirect imported input requirement in the production of $X_1$ and $X_2$ is

$$M_1^* = \sum M_i s_{i1} = a_{31}$$
$$M_2^* = \sum M_i s_{i2} = a_{32}$$

The gross labor and capital requirement per unit export is

$$\frac{\sum_j \left[ \sum_i C_{Li} s_{ij} \right] e_j}{1 - \sum_j \left[ \sum_i M_i s_{ij} \right] e_j} = \frac{C_{L1}}{1 - a_{31}}$$

$$\frac{\sum_j \left[ \sum_i C_{Ki} s_{ij} \right] e_j}{1 - \sum_j \left[ \sum_i M_i s_{ij} \right] e_j} = \frac{C_{K1}}{1 - a_{31}}$$

since $e_j$ is unity for $X_1$ and zero for $X_2$, by assumption.

Applying the gross factor-intensity measure given in equation (4) yields

$$k_1^{**} = \frac{C_{K1} + a_{31}\dfrac{C_{K1}}{1 - a_{31}}}{C_{L1} + a_{31}\dfrac{C_{L1}}{1 - a_{31}}} = \frac{C_{K1}}{C_{L1}} = k_1$$

$$k_2^{**} = \frac{C_{K2} + a_{32}\dfrac{C_{K1}}{1 - a_{31}}}{C_{L2} + a_{32}\dfrac{C_{L1}}{1 - a_{31}}} = \frac{C_{K2} - C_{K2}a_{31} + C_{K1}a_{32}}{C_{L2} - C_{L2}a_{31} + C_{L1}a_{32}}$$

$$k_3^{**} = \frac{C_{K3}}{C_{L3}} = k_3$$

Algebraic manipulation reveals that if $k_1 > k_2$, $k_1^{**}$ will always be greater than $k_2^{**}$. If the pure intermediate is totally imported, the standard trade theorems defined in net factor intensities remain valid. The intuitive reason for this result is simply that the importation of intermediate goods implicitly involves the substitution of exports (final goods) for the intermediate good in the production process—hence an interindustry flow model à la Vanek.

### C. *The Intermediate Product Partially Imported, Partially Produced Domestically*

The more general case, and certainly the most interesting empirically, is that in which some proportion of the intermediate good is imported and some proportion supplied domestically. Again it makes no difference which of the final goods are exported in exchange for the intermediate. Assume that both $X_1$ and $X_2$ are exported in exchange for $X_3$.

$$(5) \qquad k_1^{**} = \frac{[C_{K1} + C_{K3}d_{31}] + m_{31}\left[\dfrac{[C_{K1} + C_{K3}d_{31}]e_1 + [C_{K2} + C_{K3}d_{32}]e_2}{1 - m_{31}e_1 - m_{32}e_2}\right]}{[C_{L1} + C_{L3}d_{31}] + m_{31}\left[\dfrac{[C_{L1} + C_{L3}d_{31}]e_1 + [C_{L2} + C_{L3}d_{32}]e_2}{1 - m_{31}e_1 - m_{32}e_2}\right]}.$$

$$k_2^{**} = \frac{[C_{K2} + C_{K3}d_{32}] + m_{32}\left[\dfrac{[C_{K1} + C_{K3}d_{31}]e_1 + [C_{K2} + C_{K3}d_{32}]e_2}{1 - m_{31}e_1 - m_{32}e_2}\right]}{[C_{L2} + C_{L3}d_{32}] + m_{32}\left[\dfrac{[C_{L1} + C_{L3}d_{31}]e_1 + [C_{L2} + C_{L3}d_{32}]e_2}{1 - m_{31}e_1 - m_{32}e_2}\right]}$$

$$k_3^{**} = \frac{C_{K3}}{C_{L3}} = k_3$$

Given that some proportion ($\neq 1$) of the intermediate good is imported, the domestic structure of production is reflected in the inverse domestic Leontief matrix

$$[I - D]^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ d_{31} & d_{32} & 1 \end{bmatrix}$$

The proportion of $X_3$ imported and used in the production of $X_1$ and $X_2$ is denoted $m_{31}$ $(= a_{31} - d_{31})$ and $m_{32}$ $(= a_{32} - d_{32})$, respectively. Applying the gross factor-intensity measure expressed in equation (4) yields expressions for $k_1^{**}$, $k_2^{**}$, and $k_3^{**}$, shown in (5).

Defining the total capital requirement per unit export as $A$, and the total labor requirement per unit export as $B$ (for convenience since they are constants at any one point in time under a given set of prices), the condition under which $k_1^* > k_2^*$, if $k_1 > k_2$, is

$$\frac{C_{K1} + C_{K3}d_{31} + m_{31}A}{C_{L1} + C_{L3}d_{31} + m_{31}B} > \frac{C_{K2} + C_{K3}d_{32} + m_{32}A}{C_{L2} + C_{L3}d_{32} + m_{32}B}$$

or

$$\frac{k_1C_{L1} + k_3C_{L3}d_{31} + m_{31}A}{C_{L1} + C_{L3}d_{31} + m_{31}B} > \frac{k_2C_{L2} + k_3C_{L3}d_{32} + m_{32}A}{C_{L2} + C_{L3}d_{32} + m_{32}B}$$

which, provided $m_{31} = m_{32}$, holds if and only if

(i) $\qquad k_1 > k_3 > k_2$

or

(ii) $\quad k_1 > k_2 > k_3$ and $d_{32}/C_{L2} \geq d_{31}/C_{L1}$

As Batra and Casas previously demonstrated, if the pure intermediate good is produced domestically either condition (i) or (ii) is necessary in order to ensure a uniform ranking of gross and net factor intensities of final products. However, these conditions are insufficient unless the intermediate is *totally* supplied from domestic production. In the more general case in which the intermediate is both imported and supplied domestically, a further necessary condition is required, namely that the proportion imported and used in the production of the final goods is the same for both goods.

### III. Conclusion

The validity of the traditional trade theorems which depend upon the factor-intensity ranking of traded commodities demands that factor intensity be defined in gross terms, rather than in terms of more conventional net factor intensity. Empirical studies of the traditional theorems, in particular "Leontief tests" of the Heckscher-Ohlin theorem, have long recognized the necessity of defining factor intensity in gross terms. Nevertheless, what empirical studies have failed to recognize is that the conventional measure of gross intensity, introduced by Leontief and applied countless times, implicitly assumes that intermediate products are solely pro-

duced and supplied domestically. More appropriately, in an open economy importing intermediate products, gross factor intensity is measured by 1) the factor requirements at the last stage of production and in producing domestic intermediate inputs, and 2) the factor requirements of earning the foreign exchange needed to purchase imported intermediate inputs. It matters not to the relatively labor-abundant economies of Taiwan and Hong Kong, for example, that the steel which goes into transistor radio components is a relatively capital-intensive product, as long as this steel can be obtained in exchange for relatively labor-intensive commodities such as toys, cotton shirts, and the like.[10] In testing whether observed trade patterns conform to theoretical hypotheses dependent upon factor intensity ranking, this factor should be considered. As such, the measure of gross factor intensity developed in Section I is the relevant one for empirical tests of the traditional theorems, as well as for examining in a more general context the theoretical implications of introducing intermediate products into the orthodox factor-proportions trade model.

[10] This argument is developed in my 1975 article.

## REFERENCES

**R. N. Batra** and **F. R. Casas**, "Intermediate Products and the Pure Theory of International Trade: A Neo-Heckscher-Ohlin Framework," *Amer. Econ. Rev.*, June 1973, *63*, 297–311.

**M. C. Kemp**, *The Pure Theory of International Trade and Investment*, Englewood Cliffs 1964.

**W. W. Leontief**, "Domestic Production and Foreign Trade: The American Capital Position Re-examined," *Proc. Amer. Philosophical Soc.*, Sept. 1953, *97*, 332–49.

———, "Factor Proportions and the Structure of American Trade: Further Theoretical and Empirical Analysis," *Rev. Econ. Statist.*, Nov. 1956, *38*, 386–407.

**A. Ray**, "Traded and Nontraded Intermediate Inputs and Some Aspects of the Pure Theory of International Trade," *Quart. J. Econ.*, May 1975, *89*, 331–40.

**J. Riedel**, "Factor Proportions, Linkages and the Open Developing Economy," *Rev. Econ. Statist.*, Nov. 1975, *57*, 487–94.

**A. Schweinberger**, "Pure Traded Intermediate Products and the Heckscher-Ohlin Theorem," *Amer. Econ. Rev.*, Sept. 1975, *65*, 634–43.

**J. Vanek**, "Variable Factor Proportions and Interindustry Flows in the Theory of International Trade," *Quart. J. Econ.*, Feb. 1963, *77*, 129–42.

**L. L. Wegge** and **M. C. Kemp**, "Generalization of the Stolper-Samuelson and Samuelson-Rybczynski Theorems in Terms of Conditional Input-Output Coefficients," *Int. Econ. Rev.*, Oct. 1969, *10*, 414–25.

# Elasticity, Absorption, Keynesian Multiplier, Keynesian Policy, and Monetary Approaches to Devaluation Theory: A Simple Geometric Exposition

*By* HARRY G. JOHNSON*

The history of balance-of-payments theory since the early 1930's has been one of successive "approaches" of increasing degrees of theoretical sophistication. Five stages of analysis (conceptually if not always chronologically) may be distinguished: the simple "elasticity" approach following the classic paper by Joan Robinson, the "absorption" approach, the Keynesian "multiplier" approach, the Keynesian "policy" approach pioneered by James Meade, and most recently the "monetary" approach stemming from the work of Robert Mundell. Differences between these approaches have occasionally been the focus of sharp controversy, most notably in the case of the elasticity and absorption approaches, and recently in the case of the monetary approach as contrasted with other approaches that have in common an emphasis on elasticities or the influence of exchange rate changes on trade flows via relative price changes and international elasticities.

The purpose of the present note is to bring out the key differences between these alternative approaches, as exemplified by a simple case that can be illustrated by a simple diagram. The simple case is that of devaluation by a single country in a world economy so large that macro-economic repercussions of devaluation on real incomes, world money demand relative to supply, and the prices of imported goods can be ignored. A further simplification is the assumption that export supply is perfectly elastic in response to domestic currency price (cost of production is constant) short of "full employment," interpreted as a specific level of total output, after which point supply is perfectly inelastic. For simplicity, also, where the analysis involves full-employment conditions the initial equilibrium point is assumed to coincide with exact full employment. Finally, international security transactions are assumed absent, all capital movements taking the form of money flows; and all money is assumed to be international money, to avoid problems (important in reality) of substitution between international reserve assets and "domestic credit." The last assumption raises the problem that a devaluation alters the amount of domestic money valued in foreign currency, and vice versa; this problem is ignored until the end of the exposition.

Figure 1 graphs income earned from export sales $X$ plus domestic purchase of home-produced goods $cE$ against domestic expenditure $E$, both measured in domestic unit values of domestic product (at some point below it will be convenient to assume measurement in terms of foreign currency unit

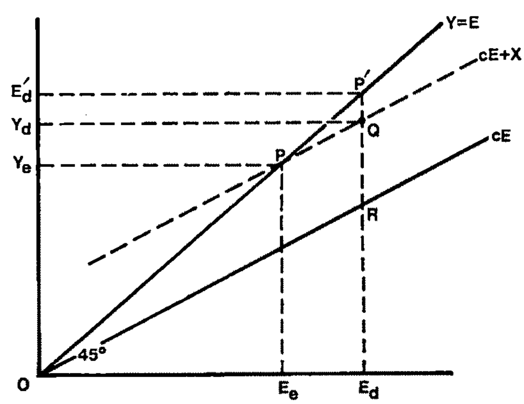



FIGURE 1

* Professor of economics, University of Chicago.

values). Export sales are an increasing function of the foreign exchange rate $r$. The division of total domestic expenditure between the vector $cE$ and the 45° line $Y=E$ is assumed to be a constant ratio relationship, and to depend on the exchange rate $r$; if the elasticity of demand for imports (with constant total expenditure) is greater than unity, $cE$ rises as $r$ rises (the exchange rate depreciates, domestic cost of exports remaining constant) and vice versa. The vector $cE+X=D$, total (foreign and domestic) expenditure on domestic output, increases or decreases with $r$ according as the well-known elasticity criterion is satisfied. This criterion can be written for the case of initial equality of exports and imports or of domestic expenditure and income as $\eta_h+\eta_f>1$, where $\eta_h$ and $\eta_f$ are the (constant-expenditures) elasticities of home and foreign demands for imports; it is worth noting that since, with expenditure constant, an excess supply of one good means an excess demand for the other, the criterion can equally well be written as $\eta_h+\epsilon_f>0$, where $\epsilon_f$ is the elasticity of foreign supply of imports, which is merely the condition that a fall in the price of a good leads to an excess demand for it.

In Figure 1, $Y_e=E_e$ at $P$ represents a situation of balance-of-payments equilibrium, with exports equal to imports and domestic expenditure equal to income; $Q$, on the other hand, is a situation of balance-of-payments disequilibrium, with imports in excess of exports and expenditure in excess of income by $P'R-QR=P'Q=E_d'Y_d$. The excess in each case is financed by a reduction of cash balances (loss of reserves) $P'Q$.

The fundamental difference between the monetary approach and the other four approaches listed consists in the fact that, according to the monetary approach, the loss of reserves and cash $P'Q$ must reduce the stock of money available next period and in subsequent periods (unless this is made good by a policy of continuous money creation by the country's monetary authorities, via substitution of domestic credit for international reserves in the portfolio of the central bank, a possibility here excluded by assumption); this will reduce the level of domestic expen-

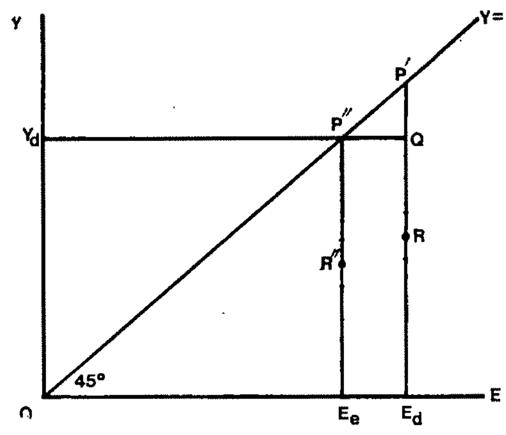



FIGURE 2

diture (to less than $E_d'$) next period; and the process will continue until equality is restored between exports and imports, and between expenditure and income, at the equilibrium point $P$ (with $Y_e=E_e$). On the other approaches, by contrast, the disequilibrium will repeat itself period after period until policy changes are introduced (in the form of devaluation, deflation, or both) that will be effective in establishing an equilibrium between exports and imports, expenditure and income, at the point $P$ or some similar point along the 45° line $Y=E$ through the origin $O$.

The difference between the other four approaches may now be examined and simply described. The "elasticity" approach, at least in one very common version of it, holds output constant (at $Y_d$ in Figure 2) and considers the conditions under which devaluation will correct the excess of imports over exports. Implicitly, this also means by assumption that devaluation corrects the excess of expenditure over income; and this involves the implicit assumption, not by any means natural to demand theory, that changes in expenditure are neutralized by changes in saving. Specifically, in diagrammatic terms the movement from disequilibrium to equilibrium (from $P'$ and $E_d$ to $P''$ and $E_e$) involves an increase in exports and reduction in imports amounting to $P'Q$ accompanied by a corresponding reduction of domestic expenditure by $E_dE_e$. The most nat-

ural assumption, though one that underlines the unreality of the necessary assumptions of the analysis, is that the proceeds of any increase in exports and the savings from any decrease in imports expenditure (noting that a smaller quantity of imports might be outweighed by a higher domestic currency price) are simply added to domestic savings or more accurately reduce domestic dissaving. Incidentally, the point $R''$ below $P''$ in the diagram, which represents the division of equilibrium income and consumption between domestic goods consumption and consumption of imports, may be above or below the level of the vector $OR$, though normally assumed to be above, according to whether the elasticity of demand for imports is assumed to be greater or less than unity; $R''$ may even be above $R$, if the elasticity is sufficiently greater than unity to offset the effect on domestic goods consumption of the reduction in total expenditure that is assumed to accompany devaluation.

The contrasting "absorption" approach, in the major variant that does not allow for the existence of unemployment and make use of the "propensity to absorb" and the multiplier, can be illustrated by assuming that the variables in Figure 2 are measured in foreign currency values, that $Y_d$ is the full-employment level of output, and that at full employment domestic currency prices of both exports and imports adjust immediately to world levels. A devaluation hence has no effect through relative price changes and elasticities, and can be effective in improving the balance of payments (in foreign currency terms) only to the extent that the inflation of domestic currency prices incidentally reduces the excess of absorption (real expenditure) over (real) income. The approach contemplates two alternative routes for such incidental absorption-reducing effects: a "Keynesian" route involving redistribution of income from workers to capitalists or taxpayers to government that may (but not necessarily will) reduce total demand as a function of aggregate real income; and a "monetary" route, involving the "real balance effect" of inflation in reducing the purchasing power of domestic money. The treatment of the real balance effect in this approach is obviously defective in recognizing the effect of inflation in reducing the real value of the nominal quantity of money through price increase while ignoring the effect of the initial deficit itself in reducing real balances by reducing nominal money stocks through payment of the deficit.

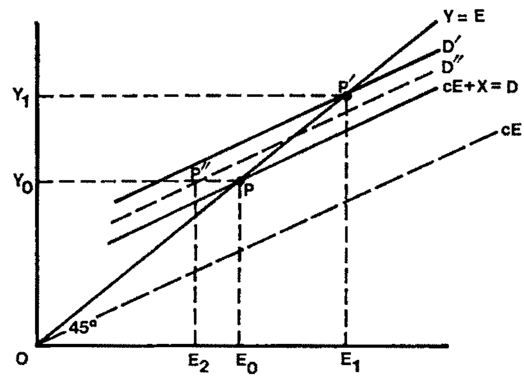



FIGURE 3

The Keynesian "multiplier" approach makes good the deficiency of the simple "elasticity" approach by recognizing and allowing for the implications of changes in expenditure on output, income, expenditure, and again output for balance-of-payments equilibrium. For simplicity, consider in Figure 3 a country initially in equilibrium with exports equal to imports and expenditure equal to income ($E_0 = Y_0$) at $P$, which proceeds to devalue in order to create a balance-of-payments surplus. The devaluation will, if the elasticity condition is fulfilled, shift the $D$ curve upward to $D'$, and, through the multiplier, increase domestic output to $Y_1$. But this fact indicates nothing about what happens to the balance of payments, even though Keynesian analysis has frequently been interpreted as showing merely that the balance of payments will improve by something less than the "primary" effect of devaluation on export and import demand. The balance of payments will improve only if the multiplier effect of higher output on total domestic expenditure (domestic goods plus imports) is less than the increase in income itself, that is, if the new $E$ lies to the

left of $E_1$ (and of course, by the usual assumption that increased income leads to increased expenditure, right of $E_0$); otherwise, the devaluation, by increasing imports more than exports, will worsen the balance of payments. The crucial requirement, given the fulfillment of the elasticity condition, is that expenditures rise less than output: in other words, that the country have a positive marginal propensity to save, or a marginal propensity to "absorb" less than unity.

Actually, it should be noted, the use of the term "marginal propensity to save" in this context is a potentially seriously misleading misnomer. For the "saving" in question is really "hoarding," the devotion of an excess of total income over total expenditure, or saving plus consumption over investment plus consumption, or of domestic saving over domestic investment, to the acquisition of additional (international) money balances. And while the term hoarding has occasionally been used in the literature, it is typically not recognized that while a positive rate of saving (acquisition of capital assets) or even (to abandon the assumption excluding international securities transactions for the moment) of net acquisition of foreign securities, may be reasonably assumed, a positive rate of acquisition of international reserves (money) related to income and persisting as a flow over all succeeding periods is not a reasonable assumption.

The "monetary" approach, by contrast, maintains that any such international monetary flow is a transient element in a stock-flow adjustment process, which must eventually come to an end. This in turn implies that the level of expenditure, whether it lies initially to the left or the right of $E_1$ (representing respectively a balance-of-payments surplus and deficit) must converge on the point $E_1$, with domestic expenditure equal to income (which has risen due to devaluation to $Y_1$, the ultimate equilibrium being depicted by $P_1$).

The Keynesian "multiplier" approach assumes unemployed resources whose rate of utilization can vary without ulterior consequences as a result of devaluation. The more sophisticated Keynesian "policy" analysis originated by Meade assumes, in contrast to previous "positive" analysis of the effects of an arbitrarily introduced devaluation, that the country under analysis has a policy authority which utilizes "financial" (fiscal and monetary) and "exchange rate" (or, in a fuller analysis, trade and exchange control) policy in order to implement objectives with respect to full employment ("internal balance") and the balance of payments ("external balance"). The analysis can be illustrated by reference again to Figure 3, where $Y_0$ now represents full-employment output (and price stability) and the balance of payments is initially in balance consistently with full employment at the point $P$ and the exchange rate and financial policy being appropriately selected to make the sums of both exports and domestic consumption, and imports and domestic consumption, equal to full-employment output $Y_0$ and expenditure $E_0$. The country now seeks to develop a balance-of-payment's surplus while maintaining full employment. If it were to deflate domestic expenditure, assuming a positive marginal propensity to spend, it would improve its balance of payments at the expense of unwanted domestic unemployment; if it devalued, assuming satisfaction of the elasticity condition, it would shift the $D$ curve upwards, requiring output above the full-employment target level (whether the increased demand would actually be satisfied, or simply produce inflationary pressure, is a question not necessary to pursue here). The solution is to combine a devaluation with a deflation in exactly the right proportions to maintain full-employment total (foreign plus domestic) demand for output while reducing total domestic demand for foreign and domestic goods below the level of total domestic output by fiscal on monetary restraint. Diagrammatically, the demand curve for domestic output can be thought of as being shifted upward by devaluation from $D$ to $D'$, and then partially shifted back downward by domestic deflation to a location $D''$, such that domestic output remains at $Y_0$, domestic expenditure is at $E_2$, and the balance-of-payments surplus and excess of domestic income over domestic expenditure is $E_2E_0$.

The Keynesian policy approach, like the other three approaches, is as mentioned subject to the criticism that it ignores the stock-flow adjustment consequences of the reserve flow associated with a balance-of-payments deficit or surplus. This adjustment process means that the combined policies of devaluation and deflation cannot produce a flow equilibrium balance-of-payments surplus of $P''D=E_2E_0$. Instead, the period-by-period expansion of cash balances consequent on the balance-of-payments surplus will gradually raise the level of expenditure; but any point on $D''$ above the level of $Y_0$ necessarily involves an inflationary excess of demand above the full-employment level, which (on a strict interpretation of "full employment") requires a domestic price level increase sufficient to reduce exports and increase imports sufficiently to make room for the additional domestic demand. Such an increase is equivalent to an appreciation of the foreign exchange rate, which partially offsets the initial devaluation; and in the long run the economy must return to the equilibrium position $P$ with no "effective" devaluation of the exchange rate (that is, the nominal devaluation is exactly offset by an inflation of domestic prices).

In conclusion, something must be said about the abstraction of the analysis from the change in the conversion ratio between domestic and foreign currencies inherent in a devaluation. This is not a problem with the "absorption" approach, since the devaluation is assumed to be offset by a corresponding domestic price inflation; and it can be ignored with the simple elasticity approach, given the prior unacceptability of the assumption that the effects of changes in export earnings and import expenditures on expenditure or domestic output are somehow neutralized. Problems arise with the Keynesian "multiplier" and "policy" models, however, inasmuch as these models assume either a varying or constant (policy maintained) level of domestic output at given domestic currency prices, and therefore a change in the foreign currency value of domestic output as a result of devaluation. In the policy case certainly, with constant value of output in domestic currency, and in the multiplier case with sufficiently low values of the relevant demand elasticities and high value of the marginal propensity to hoard, the effect of devaluation is to reduce the foreign currency value of domestic income and therefore, presumably, the country's demand for money expressed in international units. It will thereby induce a balance-of-payments *deficit* and *reserve outflow* rather than, as usually assumed, a surplus and reserve inflow, to the extent that the domestic monetary circulation consists of or is backed by international money and is therefore increased in domestic currency value by devaluation. This theoretical complication, which might have some practical implications for actual devaluation experience, can be removed by assuming that international reserves are held by an exchange equalization authority that finances its holdings by the issue of domestic-denominated liabilities and sterilizes any capital gain (or possibly loss) on existing stocks for international reserves due to devaluation (or revaluation).

## REFERENCES


S. S. Alexander, "The Effects of a Devaluation on a Trade Balance," *Int. Monet. Fund Staff Pap.*, Apr. 1952, *2*, 263–78.

J. Frenkel and H. G. Johnson, *The Monetary Approach to the Balance of Payments*, London and Toronto 1976.

J. E. Meade, *The Theory of International Economic Policy*, Vol. I: *The Balance of Payments*, London 1951.

J. Robinson, "The Foreign Exchanges," ch. 1 in *Essays in the Theory of Employment*, London 1937; reprinted in H. S. Ellis and L. A. Metzler, eds., *Readings in the Theory of International Trade*, Philadelphia 1949, 83–103.

# IN MEMORIAM

## GEORGE W. STOCKING

## 1892–1975

George Ward Stocking was born in Clarendon, Texas, on September 24, 1892, and died in Portola Valley, California, on June 7, 1975.

Clarendon was and is a small, isolated town on the dusty plain of the Texas Panhandle, where the Stocking family produced a numerous and remarkable group of brothers and sisters who had distinguished careers in government, medicine, and academic life. George graduated from Clarendon College in 1910. After following several trades, he served on the Bureau of Education of the Philippine Islands in 1915–16 and as superintendent of the Clarendon public schools in 1918–19, meanwhile finding time to get a baccalaureate degree at the University of Texas and to serve in the Air Service in World War I.

After returning to the University of Texas for an M.A. in 1921, Stocking went to Columbia for the Ph.D. in economics (1925), writing his dissertation on "The Oil Industry and the Competitive System," which remained one of his leading interests all his life. After service at the University of Vermont, the National Industrial Conference Board, and Dartmouth College, he again returned to the University of Texas at Austin in 1925. Thereafter he alternated periods of teaching with leaves of absence in New York and Washington until 1947, when he began the final phase of his academic career at Vanderbilt University in Nashville, Tennessee.

Stocking's government work included service as member and chairman of the Petroleum Labor Policy Board of the National Recovery Administration, director of the Fuels Division of OPA, and member of the War Labor Board for the 8th Region. He was a director of the Federal Reserve Branch of Dallas, San Antonio Branch, during 1943–46. For many years he served on various occasions as an arbitrator in labor disputes.

While at the University of Texas, Stocking published books on the oil industry (1925) and the potash industry (1931), and several articles on the oil industry and public policy toward it. In 1944 he began his fruitful association with the Twentieth Century Fund, in collaboration with Myron Watkins, which resulted in publication of the definitive trilogy, *Cartels in Action* (1946), *Cartels or Competition?* (1948), and *Monopoly and Free Enterprise* (1951).

As his early life might have led one to predict, George Stocking had some elements of the prairie populist in his makeup. In the unorthodox atmosphere of the University of Texas, dominated by institutional economics, Stocking was generally regarded as a conservative, even a traditionalist; in the then-ironbound conservatism of Vanderbilt he appeared as one of the first "liberals" of the postwar development. All his life he distrusted bigness and power. His major writings analyzed the origins and effects of size and power, of manipulation and conspiracy; at the same time they affirmed his hopes for the American competitive system and his belief that what had gone wrong with free competitive enterprise could be set right. At times he expressed frustration that businessmen and fellow economists did not all share his views; one of his major professional addresses was entitled "Saving Free Enterprise from its Friends." In his own work he had little taste for theoretical abstractions and legal subtleties, and showed marked antipathy toward such concepts as "workable competition" and the Rule of Reason, which he regarded as subterfuges and evasions: *genuine* competition was the goal, and any economist or lawyer should be able to recognize what it was. His published views on the matter were collected in *Workable Competition and Antitrust Policy* (Vanderbilt University Press, 1961).

The creation of the department of economics at Vanderbilt was one of his proudest achievements. He brought together a number of economists, established or promising, who took a considerable risk in going to an unproven institution in the South; it was primarily trust in George Stocking that led them to do so. He was able to attract financial support for his faculty's research, again primarily on the strength of his reputation. His deep personal concern for members of his faculty was one of his most noteworthy qualities. "The old labor arbitrator and one-time Bible salesman," one of his former associates said recently, "had a remarkable ability to get people to work together." He was a stimulating teacher who loved to play Devil's advocate to provoke debate in the classroom. He took great pains with his students and spent his time liberally with them, among other things in improving their ability to express their ideas clearly and briefly. Several of his students at Vanderbilt are now among the leading figures in industrial organization and antitrust policy.

While at Vanderbilt, George Stocking became President of the Southern Economic Association (1952) and the American Economic Association (1958). For the American Economic Association he edited, with Richard B. Heflebower, the volume of *Readings in Industrial Organization and Public Policy* (1958). Toward the end of his active career he turned his attention to the international oil industry. While spending 1960–61 in the Middle East as visiting professor at the American University at Beirut, and during other visits to the Middle East, he did extensive field research for his last book, *Middle East Oil: A Study in Political and Economic Controversy*, a timely and prophetic study published in 1970. He maintained an active interest in world oil developments to the day of his death.

# NOTES

## EIGHTY-NINTH ANNUAL MEETING OF THE AMERICAN ECONOMIC ASSOCIATION

Atlantic City, New Jersey, September 15–18, 1976

*Preliminary Announcement of the Program*

**Wednesday, September 15, 1976**

10:00 A.M. EXECUTIVE COMMITTEE MEETING

**Thursday, September 16, 1976**

8:30 A.M. EQUILIBRIUM IN MARKETS WHERE PRICE EXCEEDS COST
*Chair:* ROBERT E. HALL, Massachusetts Institute of Technology
*Papers:* DENNIS CARLTON, University of Chicago
Uncertainty, Production Lags, and Pricing
RICHARD GILBERT, University of California, Berkeley
Resource Extraction with Differential Information
A. MICHAEL SPENCE, Harvard University
Product Differentiation and Barriers to Entry
*Discussants:* PAUL JOSKOW, Massachusetts Institute of Technology
F. M. SCHERER, Northwestern University

8:30 A.M. ENVIRONMENTAL PRIORITIES: HEALTH VS. AMENITY (Joint Session with the American Association of Environmental Economists)
*Chair:* WILLIAM BAUMOL, New York University and Princeton University
*Papers:* LESTER LAVE, Carnegie-Mellon University
Evaluation of Benefits—Safety vs. Amenity
ROBERT DORFMAN AND NANCY DORFMAN, Harvard University
Socioeconomic Groups and Environmental Priorities
ALLEN KNEESE, University of New Mexico
(To be announced)
*Discussants:* (To be announced)

8:30 A.M. MICRO-ECONOMICS (Contributed Paper Session)
*Chair:* CAROLYN SHAW BELL, Wellesley College
*Papers:* JAMES B. HERENDEEN, Pennsylvania State University, AND RAYMOND LOMBRA, Board of Governors of the Federal Reserve System
A Financial Model of the Banking Firm
CAROL C. MCDONOUGH, University of Lowell
The Demand for Manuscript Review
GORDON C. WINSTON, Williams College
Capacity at the Level of the Firm, the Sector, and the Economy
S. Y. WU, University of Iowa
Monopoly Power and Stable Price Policy
*Discussant:* OLIVER WILLIAMSON, University of Pennsylvania

8:30 A.M. INCOME MAINTENANCE (Contributed Paper Session)
*Chair:* ROBERT J. LAMPMAN, University of Wisconsin, Madison
*Papers:* BENNETT HARRISON AND MARTIN REIN, Massachusetts Institute of Technology
Some Microeconomic Relationships Between Work and Welfare
KENNETH C. KEHRER et al., Mathematica Policy Research
The Initial Labor Supply Findings from the Gary Income Maintenance Experiment
ROBERT I. LERMAN, University of Wisconsin
A Reappraisal of Negative Tax and Employment Subsidy Approaches to Reforming Welfare
REBECCA A. MAYNARD, Mathematica Policy Research
The Effects of an Income Maintenance Program on School Performance of Children

Elizabeth A. Roistacher, Queens College
Short-Run Housing Responses to Income Changes
*Discussant:* Harold Watts, Columbia University

8:30 A.M. Non-Pecuniary Economics I (Contributed Paper Session)
*Chair:* Joseph L. Seneca, Rutgers University
*Papers:* Brian Forst and Kathleen B. Brosi, Institute for Law and Social Research
A Theoretical and Empirical Analysis of the Prosecutor
Itzhak Goldberg, Hoover Institution
Enforcement of Work Discipline—Detection, Rewards and Fines
Thomas J. Kniesner and Solomon W. Polachek, University of North Carolina, Chapel Hill
Scholastic Achievement: The Student as a Multi-Product Firm
Charlotte D. Phelps, Temple University
A Comparison of Adam Smith's Comments on Non-Pecuniary Rewards for Work with Contemporary Concepts and Measures of Non-Pecuniary Rewards for Work
*Discussant:* Lester Thurow, Massachusetts Institute of Technology

10:30 A.M. Endogenous Theories of Racial Discrimination (Joint Session with the National Economic Association)
*Chair:* Alfred L. Edwards, University of Michigan
*Papers:* David H. Swinton, State University of New York, Stony Brook
Labor Force Conflict Model
Gerald Jaynes, University of Pennsylvania
Information and Screening Model
Gerry Oster, City University of New York
A Marxist Model of Discrimination
*Discussants:* A. Michael Spence, Harvard University
Kenneth J. Arrow, Harvard University
Julian Ellison, Black Economic Research Center

10:30 A.M. New Directions in the Economics of Information Transfer
*Chair:* Fritz Machlup, New York University and Princeton University
*Papers:* William Baumol, New York University and Princeton University, and Yale Braunstein, New York University
An Empirical Analysis of the Production and Distribution of Information
Jack Hirshleifer and John Riley, University of California, Los Angeles
Review of New Information Theory Applications
Fritz Machlup, New York University and Princeton University
Summary
*Discussants:* Donald W. King, Market Facts, Inc.
Stephen Kagann, The John and Mary Markle Foundation

10:30 A.M. Discrimination (Contributed Paper Session)
*Chair:* Marcus Alexis, Northwestern University
*Papers:* O. L. E. Mbatia, University of California, Irvine
The Economic Effect of Fair Employment Laws on Occupations—Progress of Black Americans 1954–1972
Ralph E. Smith, The Urban Institute
Has the Recession Been an Equal Opportunity Dis-Employer?
David L. Vinje, Pacific Lutheran University
Alternative Economic Development Strategies on U.S. Indian Reservations
Michelle J. White, University of Pennsylvania
Job Suburbanization and the Welfare of Central City Minority Groups
*Discussant:* Donald W. Katzner, University of Massachusetts

10:30 A.M. Risk and Regulation in Commercial Banking (Joint Session with the American Finance Association)
*Chair:* John R. Bunting, First Pennsylvania Corporation
*Papers:* Samuel B. Chase, Jr., Samuel B. Chase, Jr., Inc., Washington, D.C.
The Regulation of Risk: Role of the Market
Andrew F. Brimmer, Harvard University
The Management of Bank Failure: Franklin National vs. Bankhaus Herstatt
Hyman P. Minsky, Washington University
Banking in a Fragile Financial Environment

*Discussants:* JACK GUTTENTAG, University of Pennsylvania
ROBERT C. HOLLAND, Board of Governors of the Federal Reserve System
CHARLES J. ZWICK, Southeast Banking Corporation
THOMAS MAYER, University of California, Davis

10:30 A.M. ANALYSIS OF DOMESTIC INFLATION
*Chair:* GARY FROMM, National Bureau of Economic Research
*Papers:* ROBERT J. GORDON, Northwestern University
Analysis of Domestic Inflation: Theory
JACK E. TRIPLETT, U.S. Bureau of Labor Statistics
Analysis of Domestic Inflation: Measurement
JOEL POPKIN, National Bureau of Economic Research
Analysis of Domestic Inflation: Transmission
*Discussants:* CHARLES L. SCHULTZE, The Brookings Institution and University of Maryland
WILLIAM NORDHAUS, Yale University
NANCY S. BARRETT, U.S. Congressional Budget Office

12:30 P.M. JOINT LUNCHEON (With the American Finance Association)
*Chair:* ALEXANDER A. ROBICHEK, Stanford University
*Speaker:* PAUL C. VOLCKER, Federal Reserve Bank of New York

2:00 P.M. ENERGY AND RESOURCES (Contributed Paper Session)
*Chair:* ROBERT S. PINDYCK, Massachusetts Institute of Technology
*Papers:* ROBERT C. ANDERSON, Environmental Law Institute
Public Policies Toward the Use of Scrap Materials
A. BRADLEY ASKIN, RICHARD L. FARMAN, GERARD L. LAGACE, AND ARTHUR J. MALLOY, Federal Energy Administration
Energy Policy and the Economy: An Empirical Analysis
DONALD A. HANSON, Ohio State University
The Value of an Exhaustible Natural Resource: Some Simple Cases
EDWARD JOHN RAY, Ohio State University
OPEC: Monopoly Control or Countervailing Power
JULIAN L. SIMON AND JOSEPH BEN-UR, University of Illinois at Urbana-Champaign
Demand, Cost, and Prices in Oligopoly: Theory, Simulation, and Oil as an Example
*Discussant:* DALE JORGENSON, Harvard University

2:00 P.M. INTERNATIONAL ASPECTS OF INFLATION
*Chair:* BERT G. HICKMAN, Stanford University
*Papers:* KARL BRUNNER, University of Rochester, AND ALAN H. MELTZER, Carnegie-Mellon University
An International Comparison of Inflation Patterns
IRVING B. KRAVIS, University of Pennsylvania and National Bureau of Economic Research, AND ROBERT E. LIPSEY, Queens College and National Bureau of Economic Research
The Adjustment of Domestic Prices to Foreign Price Changes
J. M. PARKIN, University of Western Ontario
A "Monetarist" Analysis of the Generation and Transmission of World Inflation in the 1960's and 1970's
*Discussants:* KEITH JOHNSON, Columbia University
RONALD MCKINNON, Stanford University
ROBERT J. GORDON, Northwestern University

2:00–
4:30 P.M. WELFARE ECONOMICS (Joint Session with the Econometric Society)
*Chair:* LEONID HURWICZ, University of Minnesota
*Papers:* ABBA P. LERNER, Queens College
Marginalism, Socialist and Anti-Socialist
KENNETH J. ARROW, Harvard University
Extended Sympathy in Social Choice
STANLEY REITER, Northwestern University
Information, Incentive, and Performance in the $(\text{New})^2$ Welfare Economics
*Discussants:* ABRAM BERGSON, Harvard University
THOMAS MARSCHAK, University of California, Berkeley
JERRY KELLY, Syracuse University

2:00 P.M. DISTRIBUTION OF INCOME AND WEALTH
*Chair:* IRMA ADELMAN, University of Maryland
*Papers:* GRAHAM PYATT, IBRD and University of Warwick
International Comparisons of Inequality
E. WAYNE NAFZIGER, Kansas State University
Entrepreneurship, Social Mobility, and Income Redistribution: The South Indian Case in Comparative Perspective
CHRISTOPHER CLAGUE, University of Maryland
The Supply of Skills and the Earnings Distribution
*Discussants:* DONALD HARRIS, Stanford University
CARMEL ULLMAN CHISWICK, IBRD
SHERMAN ROBINSON, Princeton University

2:00 P.M. ECONOMIC PROBLEMS CONFRONTING HIGHER EDUCATION
*Chair:* ROBERT H. STROTZ, Northwestern University
*Papers:* WALTER ADAMS, Michigan State University
(To be announced)
EARL CHEIT, University of California, Berkeley
(To be announced)
WILLIAM BOWEN, Princeton University
(To be announced)
*Discussants:* (To be announced)

4:00 P.M. SPECIAL SESSION: ECONOMIC ISSUES OF THE CAMPAIGN '76
*Chair:* (To be announced)
*Papers:* (To be announced)

8:00 P.M. RICHARD T. ELY LECTURE
*Chair:* RICHARD A. EASTERLIN, University of Pennsylvania
*Speaker:* SIMON KUZNETS, Harvard University
Two Centuries of Economic Growth—Reflections on the U.S. Experience

## Friday, September 17, 1976

8:30 A.M. AMERICAN ECONOMIC GROWTH: IMPORTED OR INDIGENOUS? (Companion Session to Richard T. Ely Lecture)
*Chair:* RICHARD A. EASTERLIN, University of Pennsylvania
*Papers:* ROBERT E. GALLMAN, University of North Carolina, Chapel Hill
Human Capital in the First Eighty Years of the Republic: How Much Did America Owe the Rest of the World?
J. R. T. HUGHES, Northwestern University
What Difference Did the Beginning Make?
NATHAN ROSENBERG, Stanford University
American Technology: Imported or Indigenous?
*Discussants:* MOSES ABRAMOVITZ, Stanford University
STANLEY LEBERGOTT, Wesleyan University
HENRY ROSOVSKY, Harvard University

8:30 A.M. RECENT CONTROVERSIES IN MONETARY THEORY
*Chair:* LEONALL ANDERSEN, Federal Reserve Bank of St. Louis
*Papers:* ROBERT CLOWER, University of California, Los Angeles
Microfoundations of Monetary Theory
EDWIN BURMEISTER, University of Pennsylvania, AND STEPHEN J. TURNOVSKY, Australian National University
Price Expectations and Stability in a Short-Run Macroeconomic Model with Money and Other Assets
JOHN RUTLEDGE, Claremont Men's College
Irving Fisher and Autoregressive Expectations
*Discussants:* KARL BRUNNER, University of Rochester
(To be announced)

8:30 A.M. THE ECONOMICS OF THE TWO-EARNER FAMILY (Joint Session with the Committee on the Status of Women in the Economics Profession)
*Chair:* ISABEL SAWHILL, The Urban Institute

*Papers:* JUNE O'NEILL, Council of Economic Advisors
The Equity of Social Security Benefits
MYRA STROBER, Stanford University
Wives' Labor Force Behavior and Family Consumption Patterns
JANE H. LEUTHOLD, University of Illinois
The Impact of Taxes on the Work Decision of the Two-Earner Family
*Discussants:* SHIRLEY JOHNSON, Vassar College
CAROLYN SHAW BELL, Wellesley College
BENJAMIN OKNER, U.S. Congressional Budget Office

8:30 A.M. APPLICATIONS OF THE DEMAND-REVEALING PROCESS (Joint Session with the Public Choice Society)
*Chair:* GORDON TULLOCK, Virginia Polytechnic Institute and State University
*Papers:* THEODORE GROVES, Northwestern University
On the Possibility of Efficient Social Choice with Compensation
WILLIAM A. VICKREY, Columbia University
Demand Revelation: Subsidies and Lump-Sum Payments
PAUL B. BROWNING AND T. NICOLAUS TIDEMAN, Virginia Polytechnic Institute and State University
Dynamically Efficient Institutions for Pollution Control
*Discussants:* KENNETH J. ARROW, Harvard University
EDWARD H. CLARKE, Office of Management and Budget
ROBERT DORFMAN, Harvard University

8:30 A.M. INCOME DISTRIBUTION (Contributed Paper Session)
*Chair:* MANORANJAN DUTTA, Rutgers University
*Papers:* BARRY CHISWICK, Council of Economic Advisers
The Effect of Americanization on Earnings
RICHARD T. CURTIN, University of Michigan
Perceptions of Distributional Equity: Their Economic Bases and Consequences
STEPHEN FRANKLIN, The Urban Institute
Microsimulation of Income and Wealth Distributions
TIMOTHY SMEEDING, University of Utah
Annual Comprehensive Income Inequality in the United States: 1968 and 1972
EDWARD STEINBERG, U.S. Department of Commerce
Earnings Increases, 1969–74
*Discussant:* MARTIN BRONFENBRENNER, Duke University

10:30 A.M. RADICAL ECONOMICS (Contributed Paper Session)
*Chair:* BERNARD SAFFRAN, Swarthmore College
*Papers:* DAVID LAIBMAN, Brooklyn College, City University of New York
Toward a Marxian Theory of Economic Growth
DALE J. POIRIER, University of Toronto
Econometric Methodology and the Radical Political Economics Literature
BRETT A. SMITH, Knox College
Worker Participation and Productivity: Another Look at the Evidence
*Discussant:* GEORGE R. FEIWEL, University of Tennessee

10:30 A.M. TAXES AND GOVERNMENT EXPENDITURE DECISIONS AFFECTING REAL ESTATE (Joint Session with the American Real Estate and Urban Economics Association)
*Chair:* JOHN KAIN, Harvard University
*Papers:* (To be announced)

10:30 A.M. APPLICATIONS OF MICROSIMULATION METHODOLOGY
*Chair:* JODIE T. ALLEN, Mathematica, Inc.
*Papers:* RICHARD R. NELSON, Yale University, AND SIDNEY G. WINTER, University of Michigan
Simulation Modeling of Schumpeterian Competition
BARBARA R. BERGMANN, ROBERT L. BENNETT, AND HARRY H. KELEJIAN, University of Maryland
Macroeconomics through Microsimulation: How Much Does Money Matter?
GUY H. ORCUTT, Yale University, AND JAMES D. SMITH, Yale University and The Urban Institute
Contextual Causes of Death Explored with Microanalytic Simulations
GUNNAR ELIASSON, Federation of Swedish Industries
Competition and Market Processes in a Simulation Model of Swedish Industry
*Discussants:* SHERMAN ROBINSON, Princeton University
JAMES L. PIERCE, University of California, Berkeley
CHARLES C. HOLT, The Urban Institute

10:30 A.M. INVITED STUDENT PAPERS (Joint Session with Omicron Delta Epsilon)
*Chair:* LOUISE NELSON, Davidson College
*Papers:* PETER MURRELL, University of Pennsylvania
Representation of Choice in Long-Term Planning
BETTY DANIEL, University of North Carolina, Chapel Hill
The Effectiveness of Stabilization Policies in Open Economies under Conditions of Inflation and Inflationary Expectations
RUSSELL P. BOISJOLY, Indiana University
Default Risk and the Theory of the Firm
*Discussants:* (To be announced)

10:30 A.M. CAPITAL FORMATION: WHERE, WHY, AND HOW MUCH?
*Chair:* ANDREW F. BRIMMER, Harvard University
*Papers:* ROBERT EISNER, Northwestern University
Capital Shortage: Myth and Reality
MARTIN S. FELDSTEIN, Harvard University
Does the United States Save Too Little?
BEATRICE N. VACCARRA, U.S. Bureau of Economic Analysis
Some Reflections on Capital Requirements for 1980
*Discussants:* BARRY BOSWORTH, The Brookings Institution
ROBERT LINDSAY, New York University
MURRAY L. WEIDENBAUM, Washington University
BURTON G. MALKIEL, Council of Economic Advisers

10:30 A.M. MARKET AND PLAN; PLAN AND MARKET (Joint Session with the Association for Comparative Economic Systems)
*Chair:* ABRAM BERGSON, Harvard University
*Papers:* RICHARD MUSGRAVE, Harvard University
The U.S. Case
ARON KATSENELINBOIGEN, University of Pennsylvania, AND HERBERT S. LEVINE, University of Pennsylvania and Stanford Research Institute
The Soviet Case
DEBORAH D. MILENKOVITCH, Barnard College
The Yugoslav Case
*Discussants:* GORDON TULLOCK, Virginia Polytechnic Institute and State University
PAUL SWEEZY, *Monthly Review*
GARY FROMM, National Bureau of Economic Research

12:30 P.M. LUNCHEON HONORING THE 1975 NOBEL LAUREATES IN ECONOMICS (Joint with the Econometric Society)
*Chair:* LAWRENCE R. KLEIN, University of Pennsylvania
*Speaker:* KENNETH J. ARROW, Harvard University

2:00 P.M. IMPACT OF RECENT DEVELOPMENTS IN PUBLIC FINANCE THEORY ON PUBLIC POLICY DECISIONS
*Chair:* JULIUS MARGOLIS, University of Pennsylvania
*Papers:* HUBERT LEVY-LAMBERT, Société Générale, Paris, France
Pricing and Investment Decisions in the French Public Sector; Recent Theoretical Developments and Practical Application
JOSEPH STIGLITZ, Stanford University
Efficiency and Equity Considerations in the Reform of the Tax System
*Discussants:* DAVID BRADFORD, Princeton University and U.S. Treasury
MARTIN FELDSTEIN, Harvard University

2:00 P.M. MONETARY THEORY FOR OPEN ECONOMIES: THE STATE OF THE ART
*Chair:* RALPH C. BRYANT, Board of Governors of the Federal Reserve System and The Brookings Institution
*Papers:* CHARLES FREEDMAN, Bank of Canada
Microeconomic Theory for Financial Intermediaries in an Open Economy
MICHAEL ADLER AND BERNARD DUMAS, Columbia University
Microeconomic Theory for Firms in an Open Economy
DALE W. HENDERSON, Board of Governors of the Federal Reserve System
Macroeconomic Theory: Modelling the Interdependence of National Money and Capital Markets
*Discussants:* RONALD L. MCKINNON, Stanford University
DWIGHT JAFFEE, Princeton University
JÜRG NIEHANS, The Johns Hopkins University

2:00 P.M. BRITISH CAPITAL IN THE LATE NINETEENTH CENTURY: SOURCES IN BRITAIN AND MOVEMENT IN THE EMPIRE
*Chair:* WILLIAM N. PARKER, Yale University
*Papers:* LANCE E. DAVIS AND ROBERT A. HUTTENBACK, California Institute of Technology
Some Foreign and Imperial Outlets
MICHAEL EDELSTEIN, Queens University
Some Domestic Sources
*Discussants:* RAYMOND W. GOLDSMITH, Yale University
WALT W. ROSTOW, University of Texas

2:00 P.M. ECONOMIC EDUCATION
*Chair:* JOHN KUHLMAN, University of Missouri
*Papers:* RENDIGS FELS, Vanderbilt University
What Economics Is Most Important to Teach: The Hansen Committee Report
ALLEN KELLEY, Duke University
New Basic Economics Courses: The JCEE Experiments
*Discussants:* LEONARD SILK, *New York Times*
MYRON JOSEPH, Carnegie-Mellon University

2:00 P.M. OTHER THINGS EQUAL (Joint Session with the Association For Co mparative Economic Studies)
*Chair:* DAVID GRANICK, University of Wisconsin
*Papers:* JOSEPH E. CHUNG, Illinois Institute of Technology
The Economies of North and South Korea
PAUL GREGORY, University of Houston
The Economies of East and West Germany
OLDRICH KYN, Boston University
The Economies of Austria and Czechoslovakia
*Discussants:* VACLAV HOLESOVSKY, University of Massachusetts
PONG LEE, State University of New York at Albany
WOLFGANG F. STOLPER, University of Michigan

2:00 P.M. GROWTH (Contributed Paper Session)
*Chair:* HERBERT LEVINE, University of Pennsylvania
*Papers:* ROGER H. BEZDEK, Energy Research and Development Administration
Technological and Structural Change in the Postwar U.S. Economy: An Input-Output Analysis
YANNIS M. IOANNIDES, Brown University
Migration and Economic Growth: Towards an Understanding of the Postwar European Experience
DANIEL R. KAZMER, Central Intelligence Agency
Agricultural Development on the Frontier: The Case of Siberia Under Nicholas II
ROGER A. MCCAIN, Temple University
The Characteristics of Optimum Inventions: An Isotech Approach
*Discussant:* WILLIAM D. NORDHAUS, Yale University

4:00 P.M. THE INVISIBLE HAND AND OTHER MATTERS (The *Wealth of Nations* Bicentennial Session)
*Chair:* WILLIAM BAUMOL, New York University and Princeton University
*Papers:* JOSEPH SPENGLER, Duke University
Adam Smith on Human Capital
SAMUEL HOLLANDER, University of Toronto
Smith and Ricardo: Aspects of the 19th Century Legacy
PAUL A. SAMUELSON, Massachusetts Institute of Technology (Read from Chair)
A Theoretical Post Mortem Vindicating Adam Smith's System

8:00 P.M. PRESIDENTIAL ADDRESS
*Chair:* JACOB MARSCHAK, University of California, Los Angeles
*Speaker:* FRANCO MODIGLIANI, Massachusetts Institute of Technology

9:30 P.M. BUSINESS MEETING

**Saturday, September 18, 1976**

8:30 A.M. NON-PECUNIARY ECONOMICS II (Contributed Paper Session)
*Chair:* PAUL RAPPOPORT, Temple University

*Papers:* CHARLES T. HAWORTH AND JOAN G. HAWORTH, Florida State University
A Model of Family Income Determination
STANLEY D. NOLLEN, Georgetown University
The College Education of Women: Labor Market Returns and Household Returns
CHARLES WOLIN, Mathematica Policy Research
Tests of the Economic Theory of Fertility
*Discussant:* WILLIAM VICKREY, Columbia University

8:30 A.M. HISTORY OF ECONOMIC THOUGHT (Joint Session with the History of Economics Society).
*Chair:* WILLIAM D. GRAMPP, University of Illinois at Chicago Circle
*Papers:* HARRY G. JOHNSON, University of Chicago
The American Tradition in Economics
G. WARREN NUTTER, University of Virginia
Political Economy in the Period of the American Revolution
*Discussants:* PAUL J. MCNULTY, Columbia University
ABRAHAM HIRSCH, Brooklyn College, City University of New York

8:30 A.M. RACIAL PROBLEMS AND INCOME DISTRIBUTION (Joint Session with the National Economic Association)
*Chair:* HUEY J. BATTLE, Virginia State College
*Papers:* JAMES D. SMITH, Yale University and The Urban Institute
Black-White Differentials in Income and Wealth in the Washington, D.C. Area
BARBARA JONES, Clark College
Black Family Income, Trends, and Patterns
CHARLES D. WHYTE, Virginia State College
New Directions for Blacks in the Labor Market: A Prospective Look at the Third Century
*Discussants:* KARL D. GREGORY, U.S. Congressional Budget Office
BERNARD E. ANDERSON, University of Pennsylvania
BRADLEY R. SHILLER, University of Maryland

8:30 A.M. ECONOMIC ANALYSES OF MEDICAL MALPRACTICE INSURANCE (Joint Session with the Health Economics Research Organization)
*Chair:* DONALD E. YETT, University of Southern California
*Papers:* MARNIE W. MUELLER, Wesleyan University
Determinants of the Supply and Demand for Malpractice Suits
PATRICIA MUNCH, Rand Corporation
A Cross-State Analysis of Medical Malpractice Litigation and Insurance Premiums
JUDITH K. MANN, University of California, San Diego
The Queuing of Malpractice Claims: Dynamic Properties of the Filing Process
*Discussants:* NEIL KOMESAR, University of Wisconsin, Madison
RONALD L. HEILMANN, University of Wisconsin, Milwaukee
WILLIAM C. HSIAO, Harvard University

10:30 A.M. NATURAL RESOURCE AND ENVIRONMENTAL CONSTRAINTS ON GROWTH (Joint Session with the American Association of Environmental Economists)
*Chair:* JOHN V. KRUTILLA, Resources for the Future
*Papers:* WILLIAM D. NORDHAUS, Yale University
Long-Run Impact of Energy Use on Climate
GARDNER M. BROWN, University of Washington, AND BARRY FIELD, University of Massachusetts
Implications of Alternative Measures of Natural Resources Scarcity
E. J. MISHAN, London School of Economics
Criteria for Intergenerational Welfare Comparisons
*Discussants:* MANCUR OLSON, University of Maryland
V. KERRY SMITH, State University of New York at Binghamton
ANTHONY C. FISHER, University of Maryland

10:30 A.M. MACRO-ECONOMICS (Contributed Paper Session)
*Chair:* ALBERT ANDO, University of Pennsylvania
*Papers:* TIM HAZLEDINE, Agriculture Canada
Keynesian Employment Functions for Canadian Manufacturing Industries
ROBERT R. KELLER, Colorado State University
Attacking Prolonged Inflation by the Coordinated Implementation of Wage-Price Controls with Contractionary Monetary Policy
SYDNEY SMITH HICKS, Florida State University
The Response of Private Interest Rates to Monetary and Debt Management Policies

ROBERT A. SCHWARTZ AND PAUL WACHTEL, New York University
The Inflationary Implications of a Variable Pricing Policy
*Discussant:* JAMES TOBIN, Yale University

10:30 A.M. REGULATION AND MICRO-ECONOMIC POLICY (Contributed Paper Session)
*Chair:* ROGER SHERMAN, University of Virginia
*Papers:* BURTON A. ABRAMS AND RUSSELL F. SETTLE, University of Delaware
Radio, Television, and the Campaign-Spending "Arms-Race": Theory and Evidence
HENRY G. GRABOWSKI AND JOHN VERNON, Duke University
Consumer Protection Regulation in the Ethical Drug Industry
HOWARD KUNREUTHER, University of Pennsylvania
Economic Analysis of Low Probability Events
JEROLD B. MUSKIN, Drexel University, AND JOHN A. SORRENTINO, Temple University
Externalities in a Regulated Industry: The Aircraft Noise Problem
*Discussant:* F. M. SCHERER, Northwestern University

10:30 A.M. INTERNATIONAL ECONOMICS (Contributed Paper Session)
*Chair:* EDWARD HEWETT, University of Texas
*Papers:* ELHANAN HELPMAN AND ASSAF RAZIN, Tel-Aviv University
The Role of Uncertainty in the Theory of International Economics: Recent Developments
JAIME DE MELO, Agency for International Development
A General Equilibrium Approach to Estimating the Allocative Costs of Protection
JAMES G. WITTE, Indiana University, AND BARBARA HENNEBERRY, Indianapolis, Indiana
The IMF Gold Auction: Some Implications for Bretton Woods
*Discussant:* PETER B. KENEN, Princeton University

# ANNOUNCEMENTS

The eighty-ninth annual meeting of the American Economic Association will be held in Atlantic City, New Jersey, September 16–18, 1976. The Employment Registry and Center will be open from September 15 to 18.

As the annual meeting is being held at an early point in the academic year, it may not be possible to complete all job placement activity through the offices of the Employment Registry and Center. Accordingly, the American Economic Association will hold a special job placement meeting for two days, at a Midwestern location, during January 1977.

## *Resolutions for Consideration at the Annual Business Meeting*

The Executive Committee at its meeting on March 8, 1974, voted to require that, to be considered at the annual business meeting, proposed resolutions must be submitted to the Secretary at least one month in advance in writing with signatures of the proposer and the second, both of whom must be members in good standing. The Secretary will reproduce the proposed resolutions and make copies available in advance of the meeting. The next business meeting will begin at 9:30 P.M. on September 17, 1976. The deadline for proposed resolutions is, accordingly, August 17. They should be sent to the Secretary, American Economic Association, 1313 21st Avenue South, Nashville, Tennessee 37212.

Nicolaus and Associates have been appointed the official travel coordinator for the Atlantic City meetings of Allied Social Science Associations. Group air fares may be arranged from your city and for groups as small as fifteen. These fares usually run about 25 percent below regular coach fare. Some examples using current fares subject to change are:

| Round Trip: | Coach | Group |
|---|---|---|
| Boston-Atlantic City | $ 91.46 | $ 61.46 |
| Detroit-Atlantic City | $161.46 | $ 87.46 |
| San Francisco-Atlantic City | $389.46 | $310.46 |

Flights will be arranged for other cities if enough individuals show interest. Air reservation forms and general information on flights and costs are available by telephoning Nicolaus and Associates collect (800+433-2897) or by writing to 1201 N. Watson Road, Suite 118, Arlington, Texas 76011.

## *Nominations for AEA Officers*

The Electoral College on March 19 chose Jacob Marschak as nominee for President-Elect of the American Economic Association in the balloting to be held in the autumn of 1976. Other nominees (chosen by the nominating committee) are: for Vice-President (two to be elected), Robert Eisner, Albert O. Hirschman, Anne Krueger, Roy Radner; for member of the Executive Committee (two to be elected), Marcus Alexis, Samuel Bowles, Robert Lampman, Marc Nerlove.

Under a change in the bylaws as described in the *Papers and Proceedings* of this *Review*, May 1971, page 472, additional candidates may be nominated by petition, delivered to the Secretary by August 1, including signatures and addresses of not less than 6 percent of the membership of the Association for the office of President-Elect, and not less than 4 percent for each of the other offices. For the purpose of circulating petitions, address labels will be made available by the Secretary at cost.

## *1977 Nominating Committee of the AEA*

In accordance with Section IV, paragraph 2, of the bylaws of the American Economic Association as amended in 1972, President-Elect Lawrence Klein has appointed a Nominating Committee for 1977 consisting of Walter W. Heller, Chairman; Nancy S. Barrett, Huey J. Battle, David M. Gordon, Bert G. Hickman, Irving B. Kravis, and Ralph W. Pfouts. Attention of members is called to the part of the bylaw reading, "In addition to appointees chosen by the President-elect, the Committee shall include any other member of the Association nominated by petition including signature and addresses of not less than 2 percent of the members of the Association, delivered to the Secretary before December 1. No member of the Association may validly petition for more than one nominee for the Committee. The names of the committee shall be announced to the membership immediately following its appointment and the membership invited to suggest nominees for the various offices to the Committee."

The Executive Committee of the American Economic Association, at its March 1976 meeting, voted to establish an ad hoc committee to explore the problems and issues involved in the federal funding of economic research, and to advise the Executive Committee whether the Association should take a more active role in seeking research support. The ad hoc committee is to report to the next meeting of the Executive Committee, September 15, 1976, in Atlantic City.

The ad hoc committee is chaired by Stanley Lebergott, Wesleyan University, and includes the following members: Milton Friedman, University of Chicago; Gary Fromm, National Bureau of Economic Research; Zvi Griliches, Harvard University; and Robert Solow, Massachusetts Institute of Technology.

Any member of the Association who is interested in the issues and wishes his or her views to be considered by the ad hoc committee is invited to communicate promptly with the chairman or any of its members.

The U.S. Census Bureau is now working on plans for the 1980 census and decisions have to be made in the relatively near future. For example, the full content of

the basic census questionnaire must be determined by the spring of 1977 so that further preparatory steps can be accomplished successfully. Although there are many constraints on the census in terms of what and how much information can be collected and tabulated, the Bureau believes that it is very important to obtain and review the recommendations of as wide a range of users and potential users of decennial census data as possible. The Census Bureau is therefore anxious to have the ideas of members of the American Economic Association. If you have any suggestions, questions, or comments on the 1980 census, please send them to the Director, U.S. Bureau of the Census, Washington, D.C. 20233.

---

The annual meeting of the International Studies Association/South will be held at the University of Virginia, Charlottesville, late October 1976. The theme of the conference is "The New International Economic Order." Persons interested in presenting papers or taking part in the program should contact the program chairman, Professor William Tyler, department of economics, University of Florida, Gainesville, FL 32611.

---

The University of Missouri-Rolla, in cooperation with the Missouri Energy Council is sponsoring the third annual UMR/MEC Energy Conference, to be held October 12–14, 1976. The theme is "Energy Crises: An Evaluation of Our Resource Potential." Papers, which may be of an applied, theoretical, or interdisciplinary nature, should be submitted for review as soon as possible. Economists from government and industry as well as academic economists are encouraged to participate. The program chairpersons are Christopher Garbacz, Vaman Rao, and William Desvousges, Economics, University of Missouri-Rolla, Rolla, Missouri 65401.

---

The Regional Science Association announces the second Victoria Lapham Dissertation Award for research on the problem of measurement, valuation, and analysis of neighborhood effects. The award will consist of a stipend of $5,000 plus tuition and fees to enable a Ph.D. candidate to devote full time to his(her) dissertation during 1976–77. For additional information, write David E. Boyce, Secretary, Regional Science Association, 3718 Locust Walk, Philadelphia, PA 19174.

---

Old Dominion University and Norfolk State College invite participation in the third annual Conference on The Urban South, to be held February 17 and 18, 1977, on the campus of Old Dominion University. The conference is designed to bring together urban specialists from all the social sciences and from public life in a series of multidisciplinary meetings and panels. Persons interested in participating in the conference should contact Professor Carl Abbott, Department of History, Old Dominion University, Norfolk, VA 23508, or Professor Robert Wolensky, Department of Sociology Norfolk, State College, Norfolk, VA 23504, by September 15, 1976.

---

A file of back issues of the *American Economic Review* (1967–75) will be donated to a library or educational institution. Anyone interested in receiving such a donation should apply to Dexter M. Keezer, Truro, Massachusetts 02666.

---

*Notice of cancellation* of the Amos Tuck School Seminar on Problems of Regulation and Public Utilities. The announcement printed in the March issue of this *Review*, page 247, should be disregarded.

---

Economists who are *strongly* oriented toward the humanities, who use humanistic methods in their research, and who will be participating in meetings abroad that are concerned with the humanistic aspects of their discipline are eligible to apply for travel grants of the American Council of Learned Societies. Specifically, economists may be eligible if (a) they deal with the history of economic thought or economic history, and (b) if their approach is qualitative and descriptive rather than quantitative and statistical. Conferences dealing with the establishment of social policy or legislation are generally ineligible. The deadlines for applications to be received in the office of the American Economic Association are: for meetings scheduled between June and September, February 15; for meetings scheduled between October and January, June 15; for meetings scheduled between February and May, October 15. Application forms may be obtained from C. Elton Hinshaw, Secretary, American Economic Association, 1313 21st Avenue South, Nashville, Tennessee 37212.

---

## *Deaths*

James J. Breen, department of economics, Clark University, Jan. 13, 1976.

William B. Gates, Jr., professor of economics, Williams College, Dec. 22, 1975.

James A. Maxwell, department of economics, Clark University, Dec. 31, 1975.

Clara Eliot Raup, Palo Alto, California, Jan. 17, 1976.

## *Retirements*

Abraham C. Burstein, director of economic research, New York City Human Resources Administration, Apr. 1975.

## *Visiting Foreign Scholars*

Pekka Ahtiala, University of Tampere, Finland: visiting professor of economics, Northwestern University, Jan. 1, 1976.

## *Promotions*

Michael S. Anselmi: assistant professor, department of economics, geography, and management, U.S. Air Force Academy, July 1, 1975.

Ashok Bhargava: assistant professor, department of

economics, University of Wisconsin-Whitewater, fall 1975.

Howard A. Clark: assistant professor, department of economics, geography, and management, U.S. Air Force Academy, July 1, 1975.

Kenneth H. Fleming: assistant professor, department of economics, geography, and management, U.S. Air Force Academy, Jan. 1, 1976.

John C. Fletcher: assistant professor, department of economics, geography, and management, U.S. Air Force Academy, July 1, 1975.

Karl D. Gregory: professor, School of Economics and Management, Oakland University.

Robert L. Gustavson: assistant professor, department of economics, geography, and management, U.S. Air Force Academy, Jan. 1, 1976.

Kenneth M. Hammer: professor of economics, University of Wisconsin-Whitewater, fall 1975.

Robert O. Heavner: assistant professor, department of economics, geography, and management, U.S. Air Force Academy, Jan. 1, 1976.

Gregory C. Hildebrandt: assistant professor, department of economics, geography, and management, U.S. Air Force Academy, Jan. 1, 1976.

Kirk Y. K. Kim: associate professor, department of economics, University of Wisconsin-Whitewater, fall 1975.

Stuart C. Kirk: assistant professor, department of economics, geography, and management, U.S. Air Force Academy, July 1, 1975.

Leon Korobow: assistant vice president, Bank Supervision and Relations Function, Federal Reserve Bank of New York, Jan. 1, 1976.

Bruce D. Mills: assistant professor, department of economics, geography, and management, U.S. Air Force Academy, Jan. 1, 1976.

Russell R. Raney: professor of economics, University of Dayton, Sept. 1976.

William E. Straw: assistant professor, department of economics, geography, and management, U.S. Air Force Academy, Jan. 1, 1976.

Robert L. Taylor: associate professor, department of economics, geography, and management, U.S. Air Force Academy, July 1, 1975.

F. Ray Webb: assistant professor, department of economics, geography, and management, U.S. Air Force Academy, July 1, 1975.

William J. Weida: assistant professor, department of economics, geography, and management, U.S. Air Force Academy, Jan. 1, 1976.

Allen E. Wolf: associate professor, department of economics, geography, and management, U.S. Air Force Academy, July 1, 1975.

Lloyd Woodman, Jr.: associate professor, department of economics, geography, and management, U.S. Air Force Academy, Jan. 1, 1976.

Charles J. Yoos II: assistant professor, department of economics, geography, and management, U.S. Air Force Academy, July 1, 1975.

Roger J. Zoeller: assistant professor, department of economics, geography, and management, U.S. Air Force Academy, Jan. 1, 1975.

### *Administrative Appointments*

Kirk Y. K. Kim: chairman, department of economics, University of Wisconsin-Whitewater, fall 1975.

David J. Cantor: associate dean, Golden Gate University, Feb. 1, 1976.

### *Appointments*

John Achterhof: research associate, department of economics, Iowa State University, Sept. 1, 1975.

Lizardo de las Casas: research associate, department of economics, Iowa State University, Oct. 1, 1975.

J. Markham Collins: assistant professor of economics, University of Dayton, Sept. 1, 1975.

Leslie C. Denend: instructor, department of economics, geography, and management, U.S. Air Force Academy, July 1, 1975.

Robert C. Drury: University of Wisconsin-Milwaukee: instructor, department of economics, University of Wisconsin-Whitewater, fall 1975.

Linda Edwards: research associate, National Bureau of Economic Research, Dec. 1, 1975

Dana R. Katnik: instructor, department of economics, geography, and management, U.S. Air Force Academy, July 1, 1975.

Abba P. Lerner: visiting professor, department of economics, Florida State University, Jan. 1976.

Hosseine Morewedge, University of Southern California: visiting professor of petroleum economics, University of Washington, Dec. 23, 1975.

Ronald H. Rasch: instructor, department of economics, geography, and management, U.S. Air Force Academy, July 1, 1975.

Dorothy P. Rice: director, National Center for Health Statistics, U.S. Department of Health, Education, and Welfare.

Thomas P. Rothrock, University of Missouri: instructor, department of economics, University of Wisconsin-Whitewater, fall 1975.

Henry H. Stick: associate professor of finance, University of Dayton, Sept. 1, 1975.

John Strauss: assistant professor, department of economics, Iowa State University, Sept. 10, 1975.

James B. Streets: instructor, department of economics, geography, and management, U.S. Air Force Academy, Jan. 1, 1976.

### *Leaves for Special Appointments*

Pekka Ahtiala, University of Tampere, Finland: Minister of Economic Affairs of Finland, 1975.

Istvan B'Racz, Quinnipiac College: National Science Foundation Faculty Fellow, department of economics, Columbia University, Jan.-Dec. 1976.

Paul F. Rizza, Slippery Rock State College: Fulbright lecturer, Helsinki School of Economics, Jan. 1976.

Ron D. White, Wichita State University: Evaluation and Systems Studies Branch, Federal Energy Administration, Feb. 2, 1976.

### *Resignations*

Richard A. Levins, Iowa State University, Sept. 16, 1975.

## NOTE TO DEPARTMENTAL SECRETARIES AND EXECUTIVE OFFICERS

When sending information to the *Review* for inclusion in the Notes Section, please use the following style:

A. Please use the following categories:

1—Deaths
2—Retirements
3—Foreign Scholars (visiting the USA or Canada)
4—Promotions
5—Administrative Appointments
6—New Appointments
7—Leaves for Special Appointments (NOT Sabbaticals)
8—Resignations
9—Miscellaneous

B. Please give the name of the individual (SMITH, John W.), his present place of employment or enrollment: his new title (if any), and the date at which the change will occur.

C. Type each item on a separate 3x5 card and please do not send public relations releases.

D. The closing dates for each issue are as follows: *March*, November 1; *June*, February 1; *September*, May 1; *December*, August 1.

This announcement supersedes and replaces a letter which was sent annually from the managing editor's office. All items and information should be sent to the Assistant Editor, *American Economic Review*, Box Q, Brown University, Providence, Rhode Island 02912.

*Please mention* THE AMERICAN ECONOMIC REVIEW *When Writing to Advertisers*

# The American Economic Review

## ARTICLES




SEPTEMBER 1976

# THE AMERICAN ECONOMIC ASSOCIATION

Founded in 1885

• Published at George Banta Co., Inc., Menasha, Wisconsin.

• THE AMERICAN ECONOMIC REVIEW, including four quarterly numbers, the *Proceedings* of the annual meetings, the *Directory*, and Supplements, is published by the American Economic Association and is sent to all members five times a year, in March, May, June, September, and December.

Association membership dues for 1976, which include a subscription to both The *American Economic Review* and the *Journal of Economic Literature*, are as follows:

$25.00 for regular members with rank of assistant professor or lower, or with annual income of $12,000 or less;

$30.00 for regular members with rank of associate professor, or with annual income of $12,000 to $20,000;

$35.00 for regular members with rank of full professor, or with annual income above $20,000;

$12.50 for junior members (registered students). Certification must be submitted yearly.

Subscriptions (libraries, institutions, or firms) are $37.50 a year. Only subscriptions to both publications will be accepted. Single copies of either journal may be purchased from the Secretary's office, Nashville, Tennessee.

For foreign postage, add: PUAS countries, $3.20; all other countries and Canada, $3.75.

• Correspondence relating to the *Papers and Proceedings*, the *Directory*, advertising, permission to quote, business matters, subscriptions, membership and changes of address should be sent to the Secretary, C. Elton Hinshaw, 1313 21st Avenue South, Nashville, Tennessee 37212. Change of address must reach the Secretary at least six (6) weeks prior to the month of publication. The Association's publications are mailed second class.

• Second-class postage paid at Nashville, Tennessee and at additional mailing offices. Printed in U.S.A.

# THE AMERICAN ECONOMIC REVIEW

September 1976

VOLUME 66, NUMBER 4



• Manuscripts and editorial correspondence relating to the regular quarterly issue of this REVIEW should be addressed to George H. Borts, Managing Editor of THE AMERICAN ECONOMIC REVIEW, Brown University, Providence, R.I. 02912. Manuscripts should be submitted in duplicate and in acceptable form and should be no longer than 50 pages of double-spaced typescript. A submission fee must accompany each manuscript: $15 for members; $30 for nonmembers. *Style Instructions* for guidance in preparing manuscripts will be provided upon request to the editor.

• No responsibility for the views expressed by authors in this REVIEW is assumed by the editors or the publishers, The American Economic Association



## Articles

**Shorter Papers**



**Editor's Announcement:**

The first four papers of this issue were invited by the editor. Later issues will devote space to comments.

# The Inflationary Mechanism in the U.K. Economy

*By* R. J. Ball and T. Burns*



Perhaps one might be forgiven for supposing that, after thirty years of continuing inflation in the United Kingdom, a clear consensus would have emerged among economists, politicians, and men of business as to the causes of that inflation and the steps necessary for its cure. Unfortunately, this is not the case, although the disagreements are often exaggerated for political purposes and the arguments polarized in the sharpness of academic debate. The latest trend in the United Kingdom is to divide the participants into "Keynesians" and "monetarists," to do battle like mediaeval knights jousting at the tournament.

The explanation of prevailing disagreements is in itself complex. It is partly ascribable to the usual difficulties of adequately testing alternative hypotheses against limited sets of time-series data when it is not always possible to reject particular hypotheses against the statistical facts. It must also be said that differences arise that are not so much closely allied to analytical arguments, but which reflect the taste or persuasion of individuals with regard to prescriptions for policy. It is not unknown for positions to be rationalized by subsequent analysis rather than the logical development proceeding in the opposite direction.[1]

In this paper, we argue that the inflationary process in the *U.K.* economy cannot be simply fitted into either a so-called monetary or a Keynesian framework. The process is a more complicated one that cannot be described simply in terms of the behavior of the money supply on the one hand or the behavior of organized labor on the other. In agreement with the monetarists, it is important to differentiate between the inflationary process under fixed and floating exchange rate regimes, while in opposition to them, we assign a more important role to price and wage inflexibilities in the analysis of imported inflation and the process of adjustment to monetary change.

The discussion is divided into two main sections. In Section I, we attempt to clear the ground by reviewing the analytical arguments that frame the conclusions drawn from empirical data, and in Section II, apply the analysis to an examination of inflation in the *U.K.* economy over the last twenty years. Finally, we make some observations with regard to inflation policy.

## I

### A. *Do Trades Unions Cause Inflation?*

To this question, the monetarist gives a clear negative answer. In his world of flexible and adjustable labor markets, a stable (not necessarily zero) rate of inflation is associated in a closed economy with a unique or "natural" rate of unemployment.[2] The story in an open economy,

* Professor of economics, and senior lecturer in economics, respectively, London Business School.

[1] It is interesting to observe the situation in the United States in the late 1950's when industrial leaders ascribed inflation to trades unions, and trades unions were ascribing it to market forces. In the United Kingdom at the time, the reverse was true. Differences in preferred policies largely account for this intercountry difference at the time.

[2] As set out by Milton Friedman (1968).

which is discussed below, is not as simple, but the conclusion about the role of trades unions remains essentially the same. Given the natural rate of unemployment and equilibrium in the overall labor market, the stable rate of inflation will then depend on the rate of increase of the money stock per unit of output. Trades unions in this story cannot affect the rate of inflation, but they can affect the rate of unemployment that is consistent with a stable rate of inflation, by helping to determine the position of the aggregate supply curve for labor. The conclusion that emerges is that the actions of trades unions cannot affect the long-term rate of inflation, but can affect the level of employment. Or, putting the matter more dramatically, trades unions do not cause inflation, they cause unemployment. This argument manifestly depends on certain assumptions about the role of supply and demand in determining the level of aggregate real wages and certain further assumptions about the flexibility of both nominal wages and prices in the adjustment process.

The critics of this position have tended to be characterized as Keynesians with no belief in the working of market forces and who espouse a particularly crude theory of the way in which the monopoly power of trades unions affects the rate of increase in costs and subsequently prices.[3] Using this model, the Keynesians in Britain have been taken to assume that the money wage level is essentially exogenous and that no attention is paid by anyone other than the hardline monetarists to the behavior of the money supply in the inflationary process. It is possible that there are those who would adopt such a view, although an examination of the literature suggests that this is a straw man largely invented rather than reflecting any major body of thought. Notable exceptions perhaps can be found in the contribution of Albert G. Hines and in a recent article by Lord Kahn. It is, however, not shared by a large number of economists who would probably describe themselves as of a Keynesian persuasion.

A more sophisticated version of the argument and one which gained some credibility in the 1960's goes as follows. In the short run, the power of trades unions does permit them to force up the level of money wages and prices at a given level of employment. In addition, it may be agreed that no inflation is likely to be sustained for any significant period without an increase in the quantity of money, ruling out for the sake of argument any escape due to a rise in the velocity of circulation. If the authorities wish to sustain the level of employment then, following this union pressure on wages and prices, the money supply must be allowed to rise to accommodate it. If the increase in the money supply does not take place then, in the longer run, persistent inflationary pressure from trades unions must result in rising unemployment until, if the traditional Phillips curve is in operation, rising unemployment will choke off the rate of inflation. Thus, while in the monetarist world the trades unions can affect the rate of unemployment, in this world they can affect both the rate of inflation and the rate of unemployment. If monetary policy is accommodating, the conclusion is that the economy is operating on a labor standard in the sense defined by John Hicks (1955) whereas, if the authorities stand firm in the face of trade union pressure, unions can choose their rate of inflation or their unemployment rate, but not both. Under this argument, the Phillips curve does not vanish in the long run as expected, but approaches actual inflation in the sense of Friedman and Edmund Phelps. There is a long-run tradeoff between the rate of unemployment and the rate of wage change. The critical issue in this scenario turns not on a simple question about the monopoly power of unions, but on the specification and nature of the Phillips curve itself. The proponents of this

[3] The argument is put with particular force by A. A. Walters.

view might, therefore, agree with the monetarists that a tight control of the money supply is an essential part of any policy to contain inflation.

The practical consequences of such a model are, however, complicated by the view held by many that the volume of unemployment required to restrain the power of the unions is neither socially nor politically acceptable, so that other means must be found through some version of incomes policy to restrain the unions from requiring the authorities to enforce the unemployment/inflation tradeoff. In principle, this is in itself an empirically testable proposition, although in practice there have been some technical difficulties in establishing an acceptable statistical estimate of versions of the Phillips curve since the mid-1960's. Moreover, there is disagreement as to whether even the more stable unemployment/wage relationships estimated before that period can be interpreted as reflecting labor market behavior or the effects of changes in profitability rather than unemployment itself. (For example, see Ball, 1973, Hicks, 1975, and Kahn.)

In addition, there is a further aspect of the role of trades unions discussed by Ball (1973) which focuses on the wage transmission mechanism that operates to diffuse wage increases through the economy as a whole. This mechanism is discussed in Section Ic below. The upshot of this argument is that the existence and raison d'être of trades unions introduces inflexibilities into the labor market in the short run in such a way that inflationary pressure is created by the adjustments of relative prices to relative costs between different sectors, which is exacerbated in an open economy operating with a fixed exchange rate.[4]

Our own view as will be clearly demonstrated in Section II of this paper is that inflation in the United Kingdom since the war cannot be related in any simple way to the monopoly power exercised by trades unions. This is not to say that the power and strength of trades unions is irrelevant to understanding the inflationary process. It is to say that the power of trades unions and the pattern of collective bargaining introduces inflexibilities into the way in which the money wage level responds to changes in the pressure of demand and forces the authorities into difficult decisions with regard to the tradeoff between unemployment and the inflation rate.

### B. *The Role of Demand*

The role of demand and its effect on output and prices cannot be discussed independently of what assumptions are made about the exchange rate regime that is in operation. We begin with a fixed rate system operating in the context of an economy such as the United Kingdom that is substantially open.

In such an economy, demand pressure will not necessarily exert a great deal of direct pressure on prices, but will manifest itself in a material deterioration in the balance of payments. As long as the exchange rate is maintained then, in the open economy, the scope for manufacturers to pass on cost increases is limited by the behavior of world prices. In this case, the major impact of wage increases that results from excessive demand would be expected to be upon nonmanufactured prices, company profits, and ultimately unemployment, although clearly there is a limit to the extent that the balance of payments can deteriorate without putting pressure upon the exchange rate. Demand pressure and the consequent power of

[4] The argument is set out by Ball (1973), pp. 254–55. A mythology as to our views on this matter since has arisen from the fact that, for short period forecasting in the early 1970's, the money wage rate was treated as exogenous. The reasons for this are set out in Ball (1971). We have never suggested that the money wage is an institutional datum determined by trade union monopoly power as some critics seem to believe. A careful reading of Ball (1973) and its first edition in 1964 should make that clear.

trades unions also have an important role in determining the views of the authorities with regard to the exchange rate, and it has been widely assumed that, if the government's "target" rate of unemployment cannot be met in association with a satisfactory balance of payments, then the exchange rate is in disequilibrium. Hence, throughout the postwar period, there has been a presumption that the exchange rate may be used as an instrument to solve the dilemma.

However, along with others,[5] it has been argued at length by Ball, Burns, and J. S. E. Laury, that under free collective bargaining the effects of devaluation on the balance of payments are only temporary and that devaluation will be followed by an adjustment of both wages and prices to the higher sterling prices of both exports and imports. Thus, under a fixed exchange rate system, the ultimate effects of excessive demand pressure will operate on the price level at one remove, by causing the exchange rate to be depreciated, so diverting the demand pressure from the balance of payments to the domestic price level, leaving the authorities faced with the inevitable choice of devaluing the currency or sitting out the required rate of unemployment.

By common consent, the role of the money supply in this fixed exchange rate system is limited. Under this system, even the monetarists must concede that the stock of money is essentially demand rather than supply determined. Normally, the required money balances will be supplied by the authorities. The monetary approach tells us that a shortage or an excess of money balances will be provided or eliminated through the medium of a balance-of-payments surplus or deficit. Given this analysis under a fixed rate system, which has largely characterized the United Kingdom since 1945, it is surprising that correlations between money and prices for the economy established by those of a monetary persuasion have been interpreted as showing the causality flowing from money to prices, rather than the other way around for most of the period. Until mid-1972, Britain operated an essentially fixed rate system which, as discussed in the next section, hardly makes it plausible to account at one stroke for the rate of inflation in simple money supply terms. As has often been emphasized (for example, by Johnson), with a fixed exchange rate system, the key monetary variable becomes the rate of domestic credit expansion and its major role is the determination of the balance of payments and the split of the increase in the money supply between domestic and overseas money.

Under a floating exchange rate system, however, the picture changes to an extent that does not seem to have been grasped in the United Kingdom where the inflation debate has followed a rather rigid cost versus demand format for over twenty years. In principle, a floating exchange rate returns to each country the power to determine its own rate of inflation rather than being dictated to by the outside world. Under a floating regime, the rules change. Interest rates have a major impact on domestic monetary conditions; demand changes and particularly monetary expansions now make themselves felt on the declining exchange rate and the subsequent inflation, rather than on the balance of payments. Large balance-of-payments deficits may occur without the predictable currency crisis that would have accompanied even smaller deficits under a fixed rate system. As argued in the next section, this means that an empirical account of the history of the inflation in the postwar years must distinguish clearly between the prefloating and postfloating eras.

[5] Particularly proponents of the monetary theory of the balance of payments, e.g., Rudiger Dornbusch, Harry Johnson, Robert Mundell. Their ideas have been used to specify models, which in principle can be estimated, e.g., David Laidler, Peter Jonson.

### C. *Productivity Inflation*

Neither the monetarists nor the proponents of the simple Phillips curve analysis have adequately integrated into their accounts of the inflationary process the stresses introduced by differential rates of productivity growth throughout the economy, both between the fast and slow growing sectors of the economy and between traded and nontraded goods.

The general thesis is that money wages and prices exhibit strong downward rigidities for a variety of reasons, including the basic role of trades unions in negotiating money wages. Given a high degree of money wage inflexibility in a downward direction, changes in relative prices in line with changes in relative costs are associated with a rise in the general price level at any given rate of unemployment. Put another way, with downward price rigidity, inflation becomes a method of distributing the gains of increased efficiency throughout the economy.[6]

The significance of this mechanism is denied by the monetarists at the outset on the general ground that such price and wage rigidity cannot be demonstrated to exist. To quote Michal Parkin, "There is no reason to suppose why wages in the fast growing sectors should not rise less quickly than productivity with prices of the output of that sector falling. Indeed, that has been quite common across a wide range of manufactures" (p. 37). We would argue that this is untrue other than in electronics and certain high technology industries. It is indeed the case that there have been major shifts in the United Kingdom in *relative* prices since the war, but one would be hard put to substantiate that there have been widespread falls in *absolute* prices; indeed the overall rate of inflation of the prices of manufactured goods discussed in the next section makes it clear that this cannot have been the case. The resolution of the issue must largely depend on circumstances as yet not experienced, rather than from historical data analysis. Parkin goes on to argue that, "Finally, the hypothesis is quite incapable of explaining the wage explosion of the late 1960's." That is indeed true, but the argument developed later in this paper does not require this to be the case. The problem raised under this heading is only part of the total story, providing a mechanism whereby inflation is transmitted throughout the economy.

[6] See Ball (1973), pp. 289–301.

### D. *Can Rising Import Prices Cause Inflation?*

Since the deterioration of the *U.K.* balance of payments in 1973, and the inflation rate in 1973 and 1974, there has been a major debate as to how far the behavior of the balance of payments and the rate of inflation can be ascribed to external influences resulting in rising levels of import prices.

The monetarist answer is simple. Rising import prices cannot cause inflation provided that the government stabilizes the supply of money, since the rise in the price of imported goods will be offset by falls in the prices of other goods, keeping the average the same. The case is simply put by Walters who argues that for import prices to affect the rate of inflation is to deny the laws of supply and demand. Furthermore, they argue that the existence of a balance-of-payments deficit implies that the country concerned is exporting inflation, because the growth of money supply will be less than the rate of domestic credit expansion and the level of demand is higher than in the world as a whole.

To this, there are three answers. The first is that, even accepting the argument at face value, there is no mention of the time span over which this behavior of the average price level is to be distributed, particularly as the *U.K.* monetarists have gone out of their way to emphasize the

long lags in adjustment in the economic system. Thus, whatever the long-run outcome of the policy prescribed, it could still be true that, over a significant period (say, eighteen months), the rise in the price level could be directly attributed to the rise in import prices.

The second answer takes us back to the whole question of price inflexibility which, as we have seen, is rejected by the monetarists in connection with the discussion of productivity inflation. A priori appeals to price flexibility do not settle the issue as to whether rising import prices can or cannot permanently raise the price level. If money prices in general are sufficiently inflexible downwards then, other things being equal, rising import prices can raise both the levels of unemployment and the price level at the same time.

Thirdly, the monetarist argument assumes a degree of substitutability between externally and domestically produced goods that does not exist. As emphasized by J. Bispham, the extreme example of this arises in the case of the oil price. No one believes (except perhaps an extreme monetarist) that for the world as a whole to have attempted to offset the inflationary effects of the rise in the world oil price by policies of severe demand restrain would not have produced a most disastrous deepening of the depression on top of what actually took place. Once again, we are back to the question of the flexibility of wages and prices and the costs of adjustment which probably as a substantive issue divide some of us more from the general position of the monetarists than any other.[7] It may well be the case that this degree of flexibility varies from country to country and where there are strong trades unions and a dominant oligopolistic industry the problem is worse. For the United Kingdom, we believe that is a problem within the range of unemployment and growth experienced since the war.

In a discussion of the impact of import prices, it is useful to distinguish between increases in the prices of world manufactures (in which market the United Kingdom can compete) and increases in the prices of commodities (in which the United Kingdom cannot compete). In the case of an increase in the world manufactured price level, it is fair to argue that this should be accompanied by a balance-of-payments surplus and it is open to the United Kingdom to resist this by stabilizing the supply of money and allowing the exchange rate to appreciate. This policy was successfully pursued by Germany during this period. The difficulty with this strategy is that, because of time lags and short-run imperfections in the market, this will normally involve a short-run loss of competitiveness as exchange rates adjust ahead of prices. Thus, during the 1972-75 period, the German dollar export price level rose faster than the *U.K.* dollar export price level.[8] This may appear to be a reasonable price to pay, but the emphasis throughout this period in Britain has been upon the need to remain "competitive" in order to protect employment and manufacturing industry. To this extent, the monetarists are right, although it is important to recognize the short-term effect of the lags involved. However, when com-

[7] In this respect, we are inclined to side in part with Friedman (1970) as against James Tobin. Price and wage inflexibility plays an important role in our analysis. Our differences with the monetarists do not primarily stem from the specification of the liquidity preference function which we believe to have been improperly specified by Keynes. However, we would not accept Friedman's view that wage inflexibility is a deus ex machina introduced out of thin air. We take it as an observed and important fact about how the *U.K.* economy has operated.

[8] Between 1972 and the first quarter of 1975, the dollar value of world manufacturing export prices rose by 66 percent. West German manufacturing prices rose over the same period by 74 percent and the United Kingdom's by 53 percent. See *NIESR Review* published by the National Institute of Economic and Social Research, November 1975, Table 19.

modity prices rise faster than manufactured prices, the nature of the problem changes. It is not possible to revalue against an adverse movement of the terms of trade against manufactures without seriously affecting manufacturing exports. The problem is to adjust to the terms of trade movement. Given the degree of price inflexibility, it is inevitable that this will lead to an increase in the average price level even if domestic costs do not change. Moreover, there is a second set of problems involved in the reaction of wages to the terms of trade movement. In the face of powerful bargaining groups, the share of labor incomes in the economy is likely to rise substantially at the expense of profits as wages rise in reaction to the price increases. This presents the government with the kind of dilemma outlined above, with the further problem that a substantial proportion of its own expenditure is upon wages and salaries. One strategy is to allow company liquidity to decline, to insist on raising state industry prices and to raise taxes to cover any increase in its costs with the inevitable effects upon unemployment. The second is to attempt to subsidize prices, to allow the budgetary position to deteriorate, to borrow from abroad, and attempt to obtain some social control of incomes. Both strategies involve severe political risks.

## II

### A. *The Inflationary Process—Pre-1967*

Following our discussion of Section I, we now apply the conclusions to an examination of inflation in the British economy over the last twenty years. Initially, it is useful to divide the period under consideration into three. The first lasts from the end of the Korean War to 1967 and the sterling devaluation. This was a period of fixed exchange rates and a time when the world price level showed only a modest average increase. Between 1954 and 1967, the average British inflation rate, measured by the retail price index, was 3.2 percent per annum while the consumer spending deflator during the same period rose on average by 3 percent per annum. This was a faster increase than in the United States and Germany during this period (1.7 and 2.3 percent per annum, respectively), but slower than in Japan (3.7 percent per annum) and France (4.25 percent per annum).[9]

TABLE 1—AVERAGE PRICE INCREASE: 1954–67

| | |
|---|---|
| World Prices of Manufactures | +1.1[a] |
| U.K. Export Prices of Manufactures | +2.2[a] |
| U.K. Wholesale Prices of Manufactures | +2.2[a] |
| U.K. Retail Price Index | +3.2 |
| U.K. Consumer Price Index | +3.0 |

*Sources: NIESR Review*, various issues, and *Economic Trends*.
[a] Shown in percent.

Britain traded substantially in the market for manufactured goods in which competition increased steadily through the years and the increase in Britain's manufactured prices was substantially set by the world price level. The average inflation rates for this period are shown in Table 1. Over these years, the world export price for manufactured goods rose by only 1.1 percent per annum, while the *U.K.* export price of manufactures rose by 2.2 percent per annum. British wholesale prices of manufactures rose by the same amount as export prices, suggesting that export prices are representative of manufacturing prices as a whole. The purchasing power parity doctrine[10] suggests that these price levels should move together because of the high substitution rates between countries brought

[9] It is worth underlining these figures. It seems nowadays to be taken for granted that the United Kingdom has always headed the inflation league since the war. This is simply not so. With the notable exceptions of the United States and Germany as quoted, the British inflation rate was less than many, until the events of the 1970's that are discussed subsequently. The source of the figures is the *NIESR Review*, November 1975, Table 18, and previous issues.
[10] See Gustav Cassell.

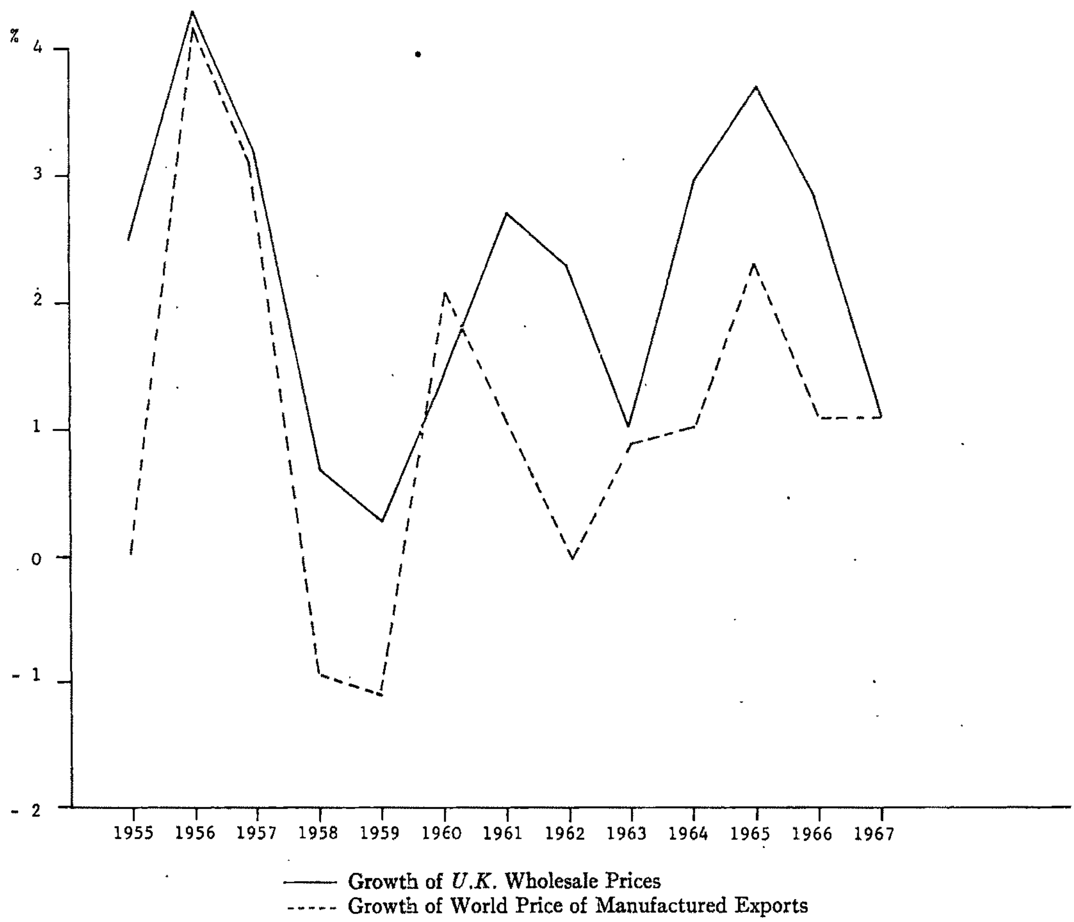



FIGURE 1. *U.K.* WHOLESALE PRICES AND WORLD MANUFACTURED EXPORT PRICES

about by the high degree of competition in the market for manufactures. Figure 1 shows that over the period[11] the cyclical movement in the growth of *U.K.* wholesale prices was substantially the same as for world export prices of manufactured goods, although throughout, British prices rose faster. Two alternative, but not incompatible, reasons can be advanced for this discrepancy. The first is that competition in the market for manufactured goods is imperfect and export prices are affected by the behavior of domestic costs which, in the United Kingdom, were rising at a faster rate than unit costs in the world as a whole. The second is that, while the equilibrium prices for manufactured goods will move together, it is possible for a country to be over- or undervalued as a result of large exchange rate changes and this discrepancy is only gradually removed. Thus, during the period of adjustment, the price level in the country that is initially undervalued will rise at a faster rate than prices in the world as a whole. Similarly, the price level in an initially overvalued currency would rise less rapidly than for the world as a whole. In principle, this process of adjustment will take place in a competitive market for demand reasons or in an

[11] See data source, Table 1.


FIGURE 2. *U.K.* AND *U.S.* REAL EXCHANGE RATE (1958 PRICES)

*Sources:* London and Cambridge Economic Service; *Economic Trends; Historical Statistics of the United States; Statistical Abstract of the United States, 1975.*

imperfectly competitive market because undervaluation of the currency reduces real income and puts upward pressure upon wages. The higher wage levels in turn lead to price increases until equilibrium is reached relative to the world price level. It seems to us quite plausible to assume that in 1954 sterling was still undervalued following the excessive devaluation of 1949 and a substantial part of the faster rise in export prices experienced in Britain represents an adjustment to the world price level as described above.

Some evidence for this view is shown in Figure 2. This shows the behavior during the century of the "real" exchange rate between Britain and the United States; i.e., the nominal exchange rate multiplied by the relative price levels in the United Kingdom and the United States. The price indices are based upon 1958 and therefore the real exchange rate is in 1958 prices. During most of this century, this real exchange rate has been relatively stable other than during wars or following substantial nominal currency appreciations and depreciations. Following these disturbances, the tendency has been for the relative price levels to adjust, so that the real average rate returns to "par," which over the period has averaged around three dollars to the pound in 1958 prices. As has been emphasized by Bela Balassa, there is no necessity for this real exchange rate to be constant; the exchange rate can only be expected in the long run to equalize traded goods prices, whereas the consumer price indices used here include nontraded prices. The movement of the consumer price index relative to export prices assuming money wage equalization within an economy will depend upon the relative growth of productivity in the traded and nontraded sectors. The faster that productivity rises in the traded sector relative to the nontraded sector, then the faster nontraded prices will rise relative to traded prices and the faster the retail price index will rise relative to export prices. A similar effect will result the higher is the share of

TABLE 2—BEHAVIOR OF WHOLESALE AND RETAIL PRICES: 1970=100

| | Prices | | Earnings | | Output per person employed | |
|---|---|---|---|---|---|---|
| | Wholesale Manufactures | Retail | Manufacturing | Whole Economy | Manufacturing | Whole Economy |
| 1963 | 78.0 | 73.9 | 64.2 | 64.9 | 77.9 | 82.1 |
| 1964 | 80.3 | 76.3 | 69.1 | 69.6 | 84.0 | 85.0 |
| 1965 | 83.3 | 80.0 | 73.8 | 74.7 | 85.4 | 87.5 |
| 1966 | 85.6 | 83.1 | 78.4 | 79.6 | 86.9 | 88.8 |
| 1967 | 86.5 | 85.2 | 81.0 | 82.1 | 90.0 | 91.7 |
| Average per annum[a] | 2.6 | 3.6 | 6 | 6 | 3.7 | 2.8 |

*Sources:* Tables 81, 90, and 93, *Economic Trends.*
[a] Shown in percent.

nontraded goods in the economy. Hence, if the "productivity gap" between the sectors in the United States was higher than in Britain, or the share of nontraded goods higher, the real exchange as measured here would fall over time. However, even if we assume the real exchange rate has been falling over the period,[12] the chart still suggests substantial undervaluation in 1950, and it is the mid-1960's before equilibrium is again reached.

Thus, to summarize the 1954–67 period, *U.K.* export prices rose in total by 15 percent more than world prices, although the cyclical movement was similar. We consider that a substantial part of this excess follows from the undervaluation of sterling at the beginning of the period although wage pressures caused in part by demand pressures may also have had a part to play.

While *U.K.* wholesale prices rose by the same amount as export prices (20 percent), retail prices on the other hand rose by an average of 3.2 percent per annum. The underlying mechanism has been outlined above. Wages in the manufacturing sector are determined by the growth of productivity in manufacturing and the increase in manufacturing prices. Demand pressures and the principle of comparability then results in wages in the nonmanufacturing sector rising by the similar magnitude. It then follows that if prices in the nonmanufacturing sector are determined by unit costs, then prices overall will rise by the growth in manufacturing prices plus a proportion (equal to the share of nonmanufacturing in the economy) of the gap between productivity growth in manufacturing and nonmanufacturing.

This process can be illustrated by the figures in Table 2 where we show the relative behavior of manufactured and retail prices. (On a consistent basis, data are only available since 1963.) Earnings in manufacturing and the whole economy rise by an almost identical 6 percent per annum, but productivity in the economy as a whole only rises by 2.8 percent per annum compared with 3.7 percent per annum for manufacturing alone.

As expected, retail prices (at 3.6 percent per annum) rise faster than manufactured prices (at 2.6 percent per annum) by an amount equal to the growth in manufacturing productivity compared to produc-

[12] The apparent downward trend is strongly influenced by the pre-1914 figures. These are particularly suspect as there are various figures for the *U.S.* consumer prices prior to 1914. For this example, the *U.S.* figures prepared by Albert Rees were used.

TABLE 3—GROWTH OF PRICES IN INDUSTRIAL COUNTRIES[a]

| Average Annual Growth 1958–1967 | Export Prices of manufactures | Consumer Prices | Consumer Prices relative to export prices |
|---|---|---|---|
| United States | 1.6 | 1.6 | 0 |
| Japan | −0.6 | 5.0 | 5.6 |
| France | 2.4 | 3.7 | 1.2 |
| Germany | 1.0 | 2.4 | 1.4 |
| United Kingdom | 1.75 | 2.75 | 0.9 |

*Sources:* Tables 18 and 19, *NIESR Review*, November 1975, and previous issues.
[a] Shown in percent.

tivity growth in the whole economy. The same principle can be seen if we compare the relative growth of total consumer prices and manufactured export prices in the large industrial countries. These figures are shown in Table 3. The results are much as the theory outlined above would suggest. Being the most rapidly growing country, Japan is well ahead of the other countries and has the largest growth of consumer prices relative to export prices; Germany comes second with France third, while Britain and the United States have the smallest gap between the growth of consumer and export prices.

Although seasonally adjusted unemployment remained below 600,000 for the entire period, the trend was unmistakably upward. Measured as a percentage of the labor force, unemployment was less than $1\frac{1}{2}$ percent at the beginning of the period and reached new heights in each recession as the undervaluation of the currency disappeared and the *U.K.* manufactured prices rose faster than world prices, while at the same time as we have already stressed, the share of profits in the economy was steadily reduced.[13] One of the features of the undervaluation of sterling in the early 1950's was the low unemployment rate that followed. The major problem was that this brought with it exaggerated views of the target level of unemployment, leading to the expectation that balance-of-payments equilibrium would be consistent in the longer run with such low rates.

As during each recession unemployment increased to yet higher levels, the authorities met this with even greater reflation in an attempt to arrest its upward movement. The combination of the reflation and the deteriorating relative price position meant that the current account deficit also increased in each cycle: from 0.9 percent of the gross domestic product (*GDP*) at the peak of the 1955 cycle to 1.1 percent in 1960 to 1.2 percent in 1964.[14] Although unemployment was high by postwar standards, the current account was in deficit at almost 1 percent of the *GDP* in 1967 making three years of deficit out of the last four. Sterling was devalued in November of that year in the mistaken belief that the inability to combine the target unemployment rate with an adequate current account performance was due entirely to an incorrect exchange rate rather than a mistaken unemployment target.[15]

## B. *The Inflationary Process: 1967–72*

The second distinct period is from 1967 to 1972. This has many of the characteristics of pre-1967, although it was now set against the background of a sharp acceleration in the world price level, the ad-

[13] Between 1955 and 1967, the share of profits, excluding stock appreciation as a proportion of total domestic incomes, fell from 16.4 to 12.9 percent.

[14] See *National Income and Expenditures*, 1964–74.

[15] This should be qualified by saying that the belief might not necessarily have been mistaken if the subsequent incomes policy pursued by the Labor government had been sustained. In the event, the "real income frustration" described by Ball (1971) persuaded the government to abandon the policy, finally allowing the real income adjustment in the face of the exchange rate change to be blown away. It was a mistaken belief to the extent that the devaluation was undertaken without a realization of the subsequent need to make the community accept the reduction in the rate of growth of real incomes.

TABLE 4—WORLD PRICES AND *U.K.* PRICES: 1967=100

| | World price manufactures | *U.K.* price of manufactures exports | | *U.K.* wholesale prices | *U.K.* retail prices | Earnings in *U.K.* | Unit labor costs |
|---|---|---|---|---|---|---|---|
| | $ | $ | £ | | | | |
| 1967 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| 1968 | 100.0 | 94.7 | 109.4 | 103.9 | 104.7 | 108.2 | 102.6 |
| 1969 | 101.1 | 97.9 | 113.0 | 108.0 | 110.3 | 116.6 | 107.4 |
| 1970 | 109.9 | 105.3 | 121.5 | 115.6 | 117.4 | 130.6 | 117.6 |
| 1971 | 115.4 | 112.6 | 131.6 | 126.0 | 128.4 | 145.4 | 128.9 |
| 1972 | 124.2 | 122.1 | 135.4 | 132.7 | 137.5 | 164.2 | 140.2 |

*Sources:* Tables 19 and 25, *NIESR Review,* November 1975, and Tables 90 and 93, *Economic Trends.*

justment of prices following the further 15 percent devaluation, and the impact of the wage and price policy in affecting the timing of these adjustments. During most of the period, the economy was subject to substantial deflation via government spending cuts and tax increases so that by 1970 the budget was in surplus to the extent of 1.7 percent of the gross domestic product compared to a deficit of 4.7 percent in 1967.[16] At the same time, this deflation combined with the exchange rate depreciation led to a substantial overseas payments surplus by 1970 and 1971, although by the end of 1971 unemployment had approached the million mark for the first time in the postwar period.

The data are set out in Table 4. After being stable in 1967–69, the world price of manufactured exports (expressed in dollars) rose by 23 percent in the following three years to a figure 24 percent above its 1967 level. The *U.K.* price of manufactured exports (measured similarly) rose by 22 percent over the same period. Naturally, following the devaluation, the British export price level fell (expressed in dollars), but by 1970 it was above its 1967 level and by 1972 it had virtually caught up with the world price and the pre-devaluation situation had been restored. Hence, expressed in sterling over the same period, export prices rose by 35 percent. Once again, the wholesale price index follows the export price level, although with an adjustment lag, but by 1972, it had risen by almost 33 percent over the 1967 level. Retail prices rose by a little under 1 percent per annum faster than wholesale prices again reflecting the productivity gap discussed above. During 1967 to 1969, when the incomes policy was in operation, unit labor costs rose by 2.6 percent less than the retail price index, but by 1970, following the collapse of the incomes policy, that gap had been closed. Wages, costs, and prices kept in step in 1971, but in 1972 wage costs moved almost 2 percent ahead of prices.

In Figure 3, we have plotted the behavior of the *U.K.* share of the world money supply on a dollar basis.[17] This falls steadily from 7 to 6.2 percent from 1963 to 1967, while the *U.K.* share of world industrial output falls from 7.5 to 6.8 percent.[18] Following the 1967 devaluation, the dollar value of the *U.K.* share of the world

[16] See *National Income and Expenditures,* 1964–74.

[17] World money supply is defined here simply as the unweighted sum of the dollar value of the money stock in the United States, Canada, Japan, France, Germany, Italy and the United Kingdom. See *International Financial Statistics.*

[18] The *U.K.*'s share of world industrial output is the ratio of the United Kingdom and O.E.C.D. Industrial Production (1970=100) multiplied by the 1970 weight of 6.2 percent. See *NIESR Review,* November 1975, Table 24.

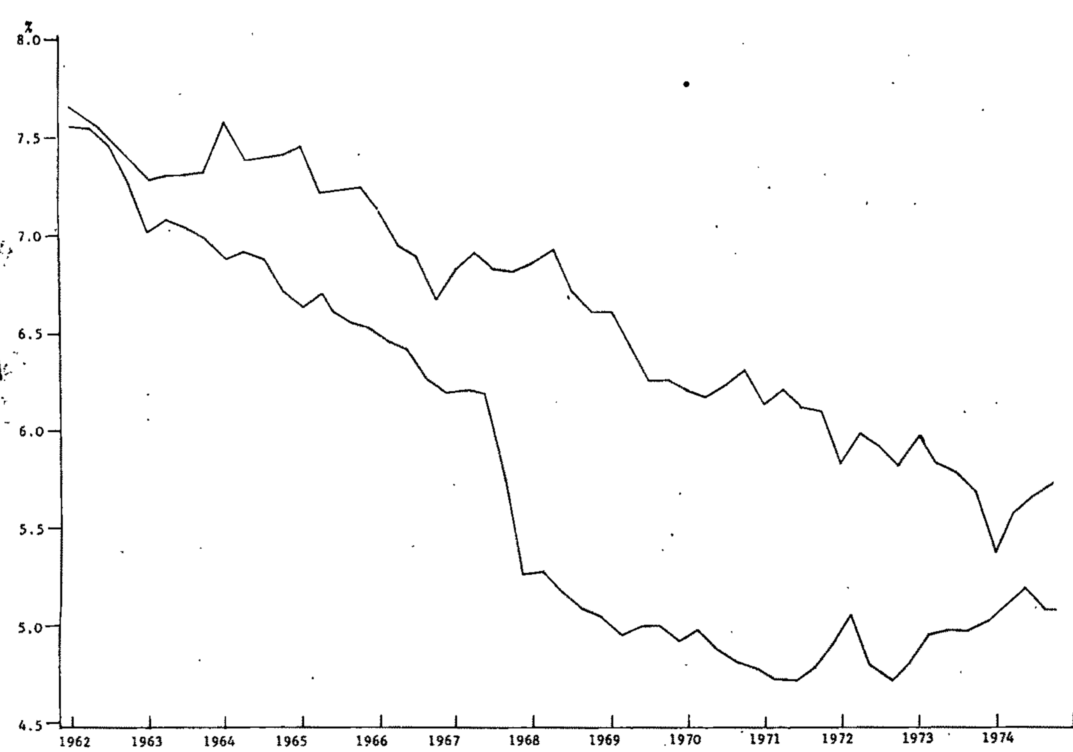



FIGURE 3. *U.K.* SHARE OF WORLD MONEY SUPPLY (BOTH IN DOLLARS) AND SHARE OF INDUSTRIAL OUTPUT

money stock fell sharply, but in the period that followed, the economy acquired the necessary money balances so that by 1972, the share was back on trend and the predevaluation situation was restored along with the price level.

We can, therefore, interpret the increased rate of inflation during the period as partly due to the higher world inflation rate and partly to the adjustment following the 1967 devaluation which in turn followed the expansionary measures of the mid-1960's. Thus, while we have argued that the demand pressures of the mid-1960's had only limited impact upon the inflation rate at the time, the inflation was merely postponed until the exchange rate was adjusted and price paid in the final years of the decade.[19]

### C. *The Inflationary Process: 1972–75*

The third distinct period lasts from 1972 to the time of writing. During this period, the world inflation rate increased still further so that the problems of domestic policy were conducted against an even more unhelpful world background. World commodity prices doubled, oil prices quintupled, and world manufactured prices (expressed in world currency)[20] rose by 35 percent between 1973 and 1975.

Since mid-1972, sterling has been allowed to fluctuate although there has still been substantial official intervention in the market. The dwindling balance-of-payments surplus that accompanied the domestic expansion triggered the crisis in July of that year. This substantially changed the situation as it effectively

[19] The postponement was also due in part to the incomes policies followed by the government from 1966–69. This further delayed the inflation that might have been expected to show itself more rapidly after the devaluation.

[20] To calculate the world price level in "world currency" we have taken the world price in dollars, transformed into sterling and multiplied by the "effective" *U.K.* exchange rate.

TABLE 5—WORLD PRICES AND *U.K.* PRICES: 1972=100

| | World price manufactures | | | Effective exchange rate | *U.K.* price of manufacturing exports | *U.K.* wholesale prices | *U.K.* retail prices | *U.K.* labor costs |
|---|---|---|---|---|---|---|---|---|
| | $ | £ | World Currency | | | | | |
| 1972 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| 1973 | 111.7 | 120.1 | 108.0 | 89.9 | 111.7 | 107.3 | 109.2 | 109.2 |
| 1974 | 143.4 | 153.4 | 132.7 | 86.5 | 141.1 | 132.4 | 126.7 | 130.0 |
| 1975 | 163.7 | 184.3 | 144.1 | 78.2 | 173.9 | 164.4 | 157.4 | 173.8 |

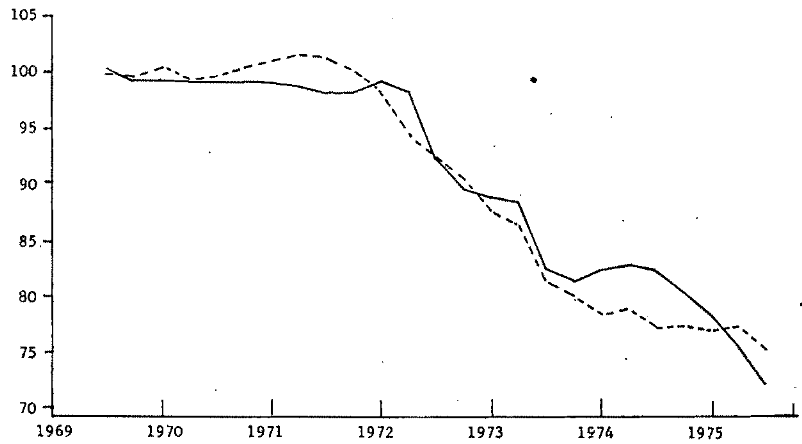
*Sources: NIESR Review*, November 1975, and *Economic Trends.*

marked the end of the fixed rate system as far as Britain was concerned. It affects our analysis as it is no longer possible to take as exogenous the sterling value of the world export price level; this, in principle, could be anything, depending upon the behavior of the exchange rate. With an endogenous exchange rate, the link with the world price level is lost and the price level within this analysis is indeterminate. Some other form is necessary to constrain its movements.

The basic information is laid out in Table 5. Between 1972 and 1975 the world price of manufactured exports, expressed this time in world currency terms, rose by 45 percent with half of the increase coming in 1974. In part, this can be attributed to the demand pressures resulting from the 1972-73 expansion, but in part it is also the result of the higher oil prices and the increased demand for manufactured exports by the oil-producing nations. In addition to this increase in world prices, the *U.K.* exchange rate relative to the world depreciated by over 20 percent, thus leading to an increase in world prices, expressed in sterling of 85 percent. British manufactured export prices and wholesale prices did not adjust fully to this rise, but by 1975 were between 65 and 75 percent higher than in 1972. Furthermore, the growth in retail prices during the past two years has been less than the growth in wholesale prices, in contrast to early years when they grew around 1 percent per annum faster. This was to be expected as the increase in oil prices changed the relative prices of manufactured goods and nonmanufactured goods in favor of manufactured goods. However, as we can see in the table, unit labor costs rose faster than wholesale prices and, therefore, the entire burden of this shift and much of the fall in output in 1975 has been felt by profits and other nonwage incomes. Between 1967 and 1972, consumer prices rose by 1 to $1\frac{1}{2}$ percent per annum faster than manufactured prices, whereas, between 1972 and 1975, the increase in consumer prices has been less than for export manufactured prices. In Germany and France, the behavior was similar with consumer prices rising less than export prices over the period and in Japan, the two prices grew by a similar amount, whereas normally consumer prices grew much faster. In the United States, the gap was much greater, probably reflecting the greater impact of reduced demand pressures.[21]

Thus, most of the difference in the behavior of inflation rates in the United Kingdom relative to the rest of the world can be seen in the behavior of the exchange rate. This raises the issue of cause and effect. Did the exchange rate adjust because of price behavior sparked off by wage demands, or did the exchange rate fall take

[21] Between 1972 and 1975 (second quarter), German export prices in local currency rose by 30 percent and consumer prices by 21 percent; in France, the figures were 46 and 35 percent; in Japan, 52 and 53 percent; in the United States, 53 and 27 percent.




——— Effective Exchange Rate (December 1971 = 100)
- - - - Ratio of World and *U.K.* Money Supply per unit of Trend Output (1970 = 100)

FIGURE 4A. EFFECTIVE EXCHANGE RATE AND MONEY SUPPLY

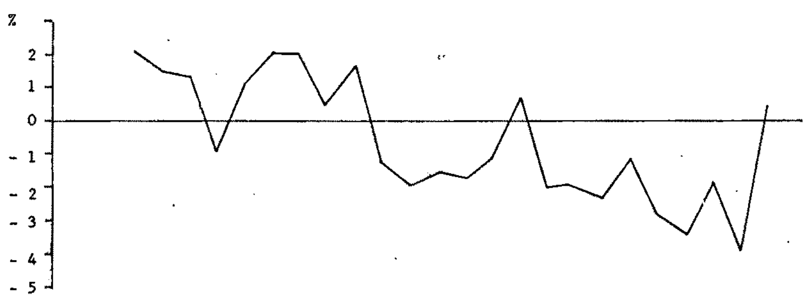



FIGURE 4B. RESERVE CHANGE AS PROPORTION OF MONEY SUPPLY

place initially because of the monetary expansion and the domestic wage and price level adjustment afterwards. In terms of timing, the evidence seems to support the latter hypothesis. The bulk of the exchange rate fall (13 percent) had taken place by September 1973 and, at that time, compared with 1972, *U.K.* prices had risen by little more than world prices.[22] The miners were making threatening noises, but the scale of the confrontation was not then evident.

In our view, there is a strong presumption that the declining exchange rate was in part related to the conduct of monetary policy. In turn this caused goods prices and factor prices to change, but in response to disequilibrium between internal and external price levels, rather than to direct demand pressures upon goods prices. Some evidence for this view is given in Figure 4. Here we have computed the ratio of the money stock to the trend of output in the United Kingdom and its main trading partners, and compared the movement of the ratio of these corrected money stock figures with the *U.K.* effective exchange rate.[23] It is clear that the money supply

[22] By the third quarter of 1973, *U.K.* consumer prices were 9.1 percent above the average for 1972. For the United States, the percentage rise was 7.3; for Germany, 7.4; for France, 8.4; for Japan, 13.5 and Italy, 11.5.

[23] Between 1967 and 1974, the ratio of United Kingdom to world industrial production fell by 2.4 percent

relative to trend industrial production in Britain started to rise much faster than for the world as a whole from the beginning of 1972. This continued until the beginning of 1974, since when the rate of increase has been slower. This is a very simple formulation of the relative excess supply of money as it assumes a unitary income elasticity in all countries. Yet the "effective" exchange rate follows the movement of the relative money variable very closely. The difference between the two lines suggests that the exchange rate was undervalued in 1971, overvalued in 1974, and back on course by the middle of 1975. This pattern is consistent with the inflow and outflow of "reserves" as measured by the difference between the increase in the money supply and the rate of domestic credit expansion as a proportion of the money supply.

The implication of this is that the United Kingdom was directly responsible for having a higher rate of inflation than the rest of the world due to inadequate control of the money supply and that, during 1974 and 1975, the internal price level was adjusting to the movement of the exchange rate in 1972 and 1973. This leaves open the question of the extent to which even the world inflation rate could have been avoided by a strict monetary policy and exchange rate appreciation.[24] As we argued above, this analysis suggests that this would indeed have been possible if the United Kingdom had been prepared to suffer the consequences of the temporary overvaluation in the same way as Germany. This issue cannot be removed from the problem of the underlying weakness of the United Kingdom growth rate. A country with an expanding share of world exports is in a much stronger position than one in which the long-run trend of unemployment has been rising and is considered politically to be "excessive."

## III

We conclude that the history of inflation in the United Kingdom since World War II reflects three forces at work of differing significance at different periods of time. The first has been the tendency to wish to run the economy from time to time at rates of growth of total demand that have not been consistent with the underlying position of the balance of payments, resulting in pressure on, and ultimately changes in, the exchange rate. The second has been the impact on the economy of external price changes, such as during the Korean War, and during the commodity boom of the 1970's culminating in the rise in the world price of oil. The third reflects the mechanism of adjustment through which efficiency increases are distributed within the economy in an institutional framework that leads to substantial downward money wage and price inflexibility.

Some unpalatable conclusions stem from this paper and from the arguments put forward in Ball, Burns, and Laury. The first is that, under a fixed exchange rate system and with free collective bargaining, it is impossible for the *U.K.* authorities to control either the rate of inflation independently from that of the rest of the world or the rate of unemployment. For any given rate of growth of world activity, there exists a rate of growth of domestic demand that is consistent with balance of payments current account equilibrium, and

per annum. This trend was used as the relative output series to deflate the ratio of the money stocks in the United Kingdom and the world in order to capture the long-run movement of output. Each series was indexed on 1970=100. The world money supply is the weighted sum of local currency money supply in the countries described above, where the weights are the 1970 exchange rate of dollars per local currency.

[24] In fact, it could be argued that it was quite unreasonable to have expected the United Kingdom to have done any better than the world average by pursuing such policies. The example of Germany hardly makes the point, since the bulk of the adjustment fell on immigrant workers who were simply shipped home. In retrospect, the criticism of the United Kingdom is that, in a sense, it chose to do worse than the world average and failed to get away with it. In the end, at least some of the unemployment came anyway. It would, however, have been very much greater, if thoroughgoing monetary prescriptions had been applied during this period.

hence, at a given moment with given technology, a certain behavior of unemployment. A rate of growth of demand above this rate can only be sustained by repeated devaluation of the currency accompanied by periodic bouts of inflation. A floating rate enables the country to regain control over the domestic price level, but does nothing to avoid the necessity of adjusting fiscal and monetary policy primarily toward the balance of payments and the exchange rate rather than the level of employment except at a major inflationary cost.

In this context, the role of an incomes policy is primarily to avoid the basic confrontation between inflation, the balance of payments, and the rate of unemployment. It is not a substitute for proper demand policies aimed at achieving an appropriate balance in government spending and taxing activities that must take the foreign balance into account. It can help in preventing the community from attempting to price itself at a level that is inconsistent with the country's overall efficiency levels in relation to world markets and to prevent erosion of the necessary rate of profitability without the necessity of creating massive unemployment in order to achieve the same objective.

Finally, it should be said that the need to restrain the demands for increases in real incomes will vary in relation to longer term issues concerning increased efficiency and industrial change in the economy that are outside the scope of this paper. The balance of payments and the rate of inflation will continue to act as constraints on increasing real incomes to the extent that fundamental problems of structure are not resolved and to the extent that explicit choices about the use of scarce resources, between private consumption and investment, and between the public and private sectors are avoided. A persistent demand for resources in excess of what can be enjoyed will only lead to further depreciation of the exchange rate, an erosion of profitability and, ultimately, to higher levels of unemployment than would otherwise have to be the case.

## REFERENCES


B. Balassa, "The Purchasing Power Parity Doctrine: A Reappraisal," *J. Polit. Econ.*, Dec. 1964, *72*, 584–96.

R. J. Ball, *Inflation and the Theory of Money*, 2d ed., London 1973.

———, "Inflation and the London Business School Model," in H. G. Johnson and A. R. Nobay, eds., *The Current Inflation*, London 1971, 43–51.

———, T. Burns, and J. S. E. Laury, "The Role of Exchange Rate Changes in Balance of Payments Adjustment—The United Kingdom Case," E.F.U. London Business School disc. pap. no. 32, July 1975.

J. Bispham, "The New Cambridge and 'Monetarist' Criticism of 'Conventional Economic Policy-Making'," *Nat. Inst. Econ. Rev.* Nov. 1975, *74*, 39–55.

G. Cassell, "The Present Situation of the Foreign Exchanges," Part I, *Econ. J.*, Mar. 1916, *26*, 62–65.

———, "Abnormal Deviations in International Exchanges," *Econ. J.*, Dec. 1918, *28*, 413–15.

R. Dornbusch, "Devaluation, Money, and Nontraded Goods," *Amer. Econ. Rev.*, Dec. 1973, *63*, 871–80.

M. Friedman, "The Role of Monetary Policy," *Amer. Econ. Rev.*, Mar. 1968, *58*, 11–17.

———, "A Theoretical Framework for Monetary Analysis," *J. Polit. Econ.*, Mar. 1971, *79*, 323–37.

J. R. Hicks, "Economic Foundations of Wage Policy," *Econ. J.*, Sept. 1955, *65*, 389–404.

———, "What is Wrong with Monetarism," *Lloyds Bank Rev.*, Oct. 1975, 1–13.

A. G. Hines, "Trade Unions and Wage Inflation in the United Kingdom, 1893–1961," *Rev. Econ. Stud.*, Oct. 1964, *31*, 221–52.

H. G. Johnson, "The Monetary Approach to Balance of Payments Theory," in *Further Readings in Monetary Economics*, London 1964, 229–49.

P. D. Jonson, "Money and Economic Activity in the United Kingdom, 1880–1970," *J. Polit. Econ.*, forthcoming.

R. F. Kahn, "Thoughts on the Behaviour of

Wages and Monetarism," *Lloyds Bank Rev.*, Jan. 1976, 1–12.

D. Laidler, *Essays on Money and Inflation*, Manchester; Chicago 1975, ch. 9.

R. A. Mundell, "Devaluation," in his *Monetary Theory: Inflation, Interest and Growth in the World Economy*, Pacific Palisades 1971, ch. 9.

M. Parkin, "United Kingdom Inflation: The Policy Alternatives," *Nat. Westminster Bank Rev.*, May 1974, 32–48.

E. S. Phelps, "Phillips Curves, Expectations of Inflation and Optimal Unemployment Over Time," *Economica*, Aug. 1967, *34*, 254–81.

A. Rees, *Real Wages in Manufacturing 1890–1914*, University Microfilms, Ann Arbor 1961.

J. Tobin, "Friedman's Theoretical Framework," *J. Polit. Econ.*, Sept. 1972, *80*, 852–64.

A. A. Walters, "Macro-economic Models and Policy in Britain," in M. D. Intriligator, ed., *Frontiers of Quantitative Economics*, Vol. III, Amsterdam 1976.

*Economic Trends*, annual supplement, Dec. 1975.

*International Financial Statistics*, various issues.

London and Cambridge Economic Service, "The British Economy Key Statistics, 1900–1970," London 1971.

*National Income and Expenditures*, 1964–1974.

National Institute of Economics and Social Research, *National Institute Economic Review (NIESR Review)*, various issues.

U.S. Bureau of the Census, *Historical Statistics of the United States, Colonial Times to 1957*, Washington 1960.

———, *Statistical Abstract of the United States 1975*, 96th ed., Washington 1975.

# Inflation in Britain: A Monetarist Perspective

*By* David Laidler*

"... the ideas which civil servants and politicians and even agitators apply to current events are not likely to be the newest. But, soon or late, it is ideas, not vested interests, which are dangerous for good or evil." *J. M. Keynes*

## I

From 1953 to 1969 the British inflation rate never exceeded 5 percent, and the unemployment rate rose above 3 percent of the labor force in only one quarter (1:1963). In 1975 the inflation rate reached a peak of over 30 percent in the second quarter, unemployment reached 5 percent of the labor force, and at the end of the year, was still rising rapidly. The contention of this paper is that Britain's difficulties in the 1970's arise from ill-designed policies based upon long and widely held misconceptions about how the economy works. The basic error committed has been to neglect to control the money supply while pursuing an unrealistically low unemployment target, primarily by fiscal means.[1] Monetary expansion, largely a by-product of "full employment" fiscal policies, has been responsible for the high British inflation rate of the early 1970's.

Figure 1 portrays the time-series behavior of inflation, unemployment, and two measures of the rate of monetary expansion over the period 1953–75.[2] To argue

* University of Western Ontario. This paper draws heavily on the work of the University of Manchester SSRC Inflation Workshop of which I, with Michael Parkin, was joint director. An earlier version of this paper was given at the University of Western Ontario Money Workshop and the University of Chicago Money Workshop. Extremely valuable criticism was received on these occasions. I am particularly grateful to Parkin, Douglas Purvis, Milton Friedman, and Robert Lucas for their comments, but nevertheless exonerate them of all responsibility for the contents of this version.

[1] The view that the quantity of money is unimportant is an article of faith in British Keynesian economics, receiving its most famous statement in the Radcliffe *Report*. It is still influential. See, for example, the evidence of Wyn Godley and Lord Kahn to the House of Commons Public Expenditure Committee in 1974. See House of Commons (1974, paras. 25–26, p. 18 and paras. 269–70, pp. 90–91, respectively).

There has, however, always existed a minority of economists in Britain who have taken the opposite position on this matter. See, for example, the evidence of Lionel (now Lord) Robbins and Frank Paish to the Radcliffe Committee, Alan Walters, and Harry Johnson. So-called "monetarist" views however have never, as far as one can tell, had any influence on policy. The reader who wishes to assess the extent to which the author of this paper is being wise after the event is referred to the author (1975, chs. 4 and 10), written in late 1970 and early 1972, respectively.

[2] Two measures of monetary expansion are presented here because it is often asserted that the two series in question tell markedly different stories. However, such differences are much more apparent on a quarter by quarter basis than they are over a longer time period. The long-term qualitative behavior of the series is very much the same. My own preference has always been to pay rather more attention to the broader money supply series, but nothing crucial hinges on this.

Note that with the introduction of the "Competition and Credit Control" reforms in late 1971 the banking system became much freer to produce near monies bearing interest at competitive rates. Such liabilities are included in $M_3$ but excluded from $M_1$. Thus the sharp downturn in the rate of expansion of $M_1$ at the beginning of 1972 reflects not a turnaround in monetary policy, but a substitution out of demand deposits into various types of time deposits suddenly available on more favorable terms. By the same token, the continued acceleration of $M_3$ growth through 1972 certainly owes something to this same effect and overstates the expansiveness of monetary policy. At the end of 1973, changes in the reserve requirements against interest bearing liabilities made it much less attractive for the banks to emit these. The rather rapid expansion of $M_1$ in 1974 at least partly reflects a shift towards non-interest bearing deposits rather than renewed monetary expansion, while the data on $M_3$ overstate the tightness of policy for similar reasons. Nevertheless, with greater flexibility of interest rates on near monies, one would expect the turning point in the $M_1$ series to come a little

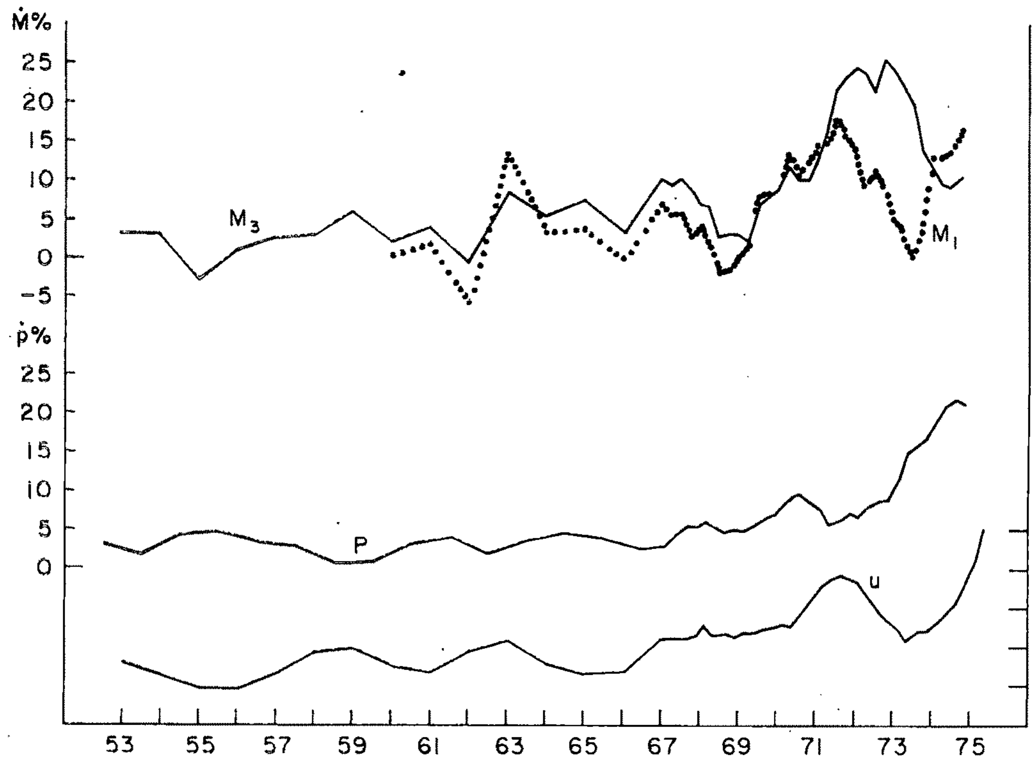



FIGURE 1

$M_1$= Annual first differences of the natural logarithm of Narrow Money; 1960–67 observations are annual, thereafter quarterly. Basic data are for end-period stocks. *Source: IFS*, various issues.

$M_3$= As $M_1$ but basic data are for money plus quasi money. *Sources:* 1960–75 *IFS*, various issues; 1953–60 *Bank of England Statistical Abstract*. Note that the Bank of England data include government deposits with Commercial Banks and the *IFS* data exclude them. This minor inconsistency between the series is probably of little importance when estimating percentage rates of change.

$P$= Annual first difference of the natural logarithm of the retail price index; 1953–67 basic data are annual averages centered in midyear; 1967–75 data are quarterly averages centered in mid quarter. *Source: NIESR Review*, various issues.

$U$= Period averages of the percentage of the labor force unemployed, Great Britain; 1953–67 annual averages; 1968–75 quarterly averages, seasonally adjusted. *Source: NIESR Review*, various issues.

that variations in the rate of monetary expansion are the main cause of variations in the inflation rate by no means implies that there should be any simple correlation between the two series, as I have shown elsewhere (1975, ch. 7). Nevertheless, results presented there for the *U.S.* economy do suggest that marked changes in the monetary expansion rate have a discernible impact on the inflation rate with about a two-year time lag, and the

before that in $M_3$ when monetary policy is changed. For example, as policy tightens up, short-term interest rates rise, and a substitution out of demand deposits into interest bearing bank liabilities involves the growth rate of $M_1$ turning down while that of $M_3$ is temporarily sustained.

Finally note that for the period over which quarterly data are plotted the monetary expansion rates are annual logarithmic first differences of end quarter data; these changes are centered on the end of the second quarter of the one-year period over which they are measured, and, as is well known, such measures tend to shift turning points backwards in time when these are sharp. These considerations are important in assessing just when, in 1973, monetary policy began to tighten up. My own inclination is to argue that the process began gently early in the year, being associated with the authorities beginning to meet a much larger fraction of government borrowing by securities sales to the nonbank public and became severe in the final quarter.

data charted in Figure 1 are consistent with the claim that a roughly similar relationship governs the behavior of British data. The pattern is particularly marked from 1968 onwards, but some might claim to discern a similar, though much less clear-cut, relationship for earlier years as well.[3] The transmission mechanism between monetary expansion and the inflation rate that underlies this interpretation of the evidence involves the initial impact of monetary changes falling on real aggregate demand—for which the unemployment rate may be regarded as a proxy—and thence influencing the inflation rate. Inspection of Figure 1 shows that such a "Phillips curve" relationship is clearly present in the pre-1967 data, vanishes between 1967 and 1971, and then begins to reestablish itself, although at much higher levels of inflation and unemployment, thereafter.

Two interrelated questions are prompted by the data charted in Figure 1. First, the pursuit of high employment with fiscal policy was just as much a characteristic of British policy before 1967 as after: why did a policy that apparently succeeded for nearly two decades fail so badly in the 1970's? Second, why did the Phillips curve disappear between 1967 and 1971? The answers offered to these questions in the next few pages may be summarized as follows. A fixed exchange rate and a low inflation rate in the world economy lay at the root of the apparent success of Keynesian policies in Britain before 1967. These policies led to the devaluation of 1967, which coincided with the beginning of the Vietnam War inflation that ultimately destroyed the Bretton Woods system. The way in which these changes impinged upon Britain accounts for the temporary disappearance of the Phillips curve after 1967. However this phenomenon was interpreted by those in control of British policy, as confirming a belief, widely held in the 1950's but temporarily unfashionable in the 1960's, that inflation was to be explained in terms of "wage push" factors, many of them noneconomic. As a direct result of this misreading of the evidence, in 1972 fiscal policy accommodated by monetary expansion and a flexible exchange rate was combined with wage and price controls in an attempt simultaneously to reduce unemployment, increase real growth, and reduce the inflation rate. The current condition of the British economy is the direct consequence of this policy, but because this diagnosis is not widely accepted in Britain, there is now a grave danger that the errors of 1972–73 might be repeated.

## II

A version of the "natural unemployment rate" hypothesis is an important implicit component of this paper's basic thesis. My first task is to show how this hypothesis can be reconciled with the behavior of the economy over the 1953–67 period. The expectations augmented Phillips curve in terms of which this hypothesis is usually formulated may be written:

[3] Nevertheless the case for attributing a role to monetary policy in generating inflation during the pre-1967 period does not rest on Figure 1. Rather it rests on the studies of Michael Artis and Bob Nobay and Bank of England (1970a, b) which use data from this earlier period. They do find evidence consistent with the view that changes in the quantity of money influenced certain variables, including money income and hence the inflation rate with a "long and variable" time lag. However, the results these studies generated were certainly much less clear-cut than those produced by similar studies done for the *U.S.* economy. Charles Goodhart of the Bank of England, the author of one of the Bank's studies, expressed a similar view of the results generated by these studies in 1974. See House of Commons (1974, para. 346, pp. 111–12). The extent to which these results reflect the fact that the British economy in the 1950's and 1960's was so stable that there was little systematic variation in income and prices to explain, and the extent to which they reflect the fact that the interaction of money and prices in a rather open fixed exchange rate economy where an interest rate stabilization policy is being pursued is more complicated than in a relatively closed economy, is too complex a question to enter into here. For a simple formal analysis of the interaction of money and inflation in a fixed exchange rate open economy model, see the author (1975, ch. 9).

$$(1) \qquad p = g(u) + X \qquad g' < 0$$

where $u$ is the deviation of unemployment from its "natural rate," $p$ is the inflation rate, and $X$ the expected inflation rate. It is common to postulate that expectations of inflation are formed according to a first-order error learning scheme applied to the actual inflation rate, but this hypothesis ignores the fact that data on the past time path of inflation are not usually the only information readily available and relevant to forecasting its future time path; where such extra information is available, economic agents are likely to make use of it. In a fixed exchange rate open economy, the time path of the price level in the rest of the world is of vital importance. It does not take much sophistication to know that if a fixed exchange rate is to be maintained, domestic inflation cannot forever deviate from that ruling in the world economy. If agents expect the exchange rate to be maintained, then they will also expect the domestic inflation rate to converge on that ruling abroad. There must be many specific hypotheses about the formation of inflationary expectations in an open economy compatible with this rather general set of propositions, but one, admittedly *ad hoc*, model that I have experimented with elsewhere with considerable success, has expectations about domestic inflation being revised according to a weighted average of the deviations of domestic and world inflation from the previously expected rate of inflation.[4] Thus, with $\Pi$ as the world rate of inflation, the change in the expected inflation rate is written as

$$(2) \quad X - X_{-1} = d[v(p_{-1} - X_{-1}) + (1-v)(\Pi_{-1} - X_{-1})] \quad 0 < d, v < 1$$

Combining this expression with equation (1) and solving for a situation in which $p = p_{-1}$ and $u = u_{-1}$ yields

$$(3) \qquad p = \frac{g}{1-v} u + \Pi_{-1}$$

Equation (3) *appears* to show that a "long-run" tradeoff between inflation and unemployment is implicit in this approach to modeling the Phillips curve for an open economy, that a lower unemployment rate may be maintained forever at the cost of a higher but not accelerating inflation rate.[5] However, this result is an illusion. It can hold only for so long as the exchange rate is held constant and is expected to remain constant. The true implication of equation (3) is that if an open economy maintains a lower than natural unemployment rate, this will result in a higher inflation rate than that ruling in the rest of the world and hence in a secularly worsening balance of payments situation. What appears at first sight to be a steady state will be one for just so long as foreign exchange reserves permit it to be sustained.

In Britain, over the period 1953–67, too low an unemployment target was set, the country experienced an inflation rate higher on average than that ruling in the rest of the world, and that higher inflation rate was associated with a secular deterioration in the balance of payments.[6] Apparently successful "Keynesian" full-employment policies thus culminated in the devaluation of November 1967. Data illustrating this argument are charted in Figure 2. Devaluation was bound to produce a

[4] See Rodney Cross and the author, and the author, forthcoming.

[5] It is worth noting that for some countries Cross and I found that $v$ was not different from zero, implying that there is no difference in slope between the long- and short-run inflation unemployment tradeoffs in those countries; Britain, however, was not one of them, $v$ taking a value of about one-half for that country.

[6] But the unemployment target was perhaps not set very far above the natural rate. Malcolm Gray, Parkin, and Michael Sumner estimate the natural rate to have been 1.8 percent over the 1953–67 period, while the actual rate averaged 1.54 percent. Their estimate is subject to a wide margin of error and certainly looks low, but is not obviously inconsistent with Phillips' estimate that a 2.5 percent unemployment rate was compatible with price stability in postwar Britain, since the expected inflation rate was surely slightly positive over the relevant period.

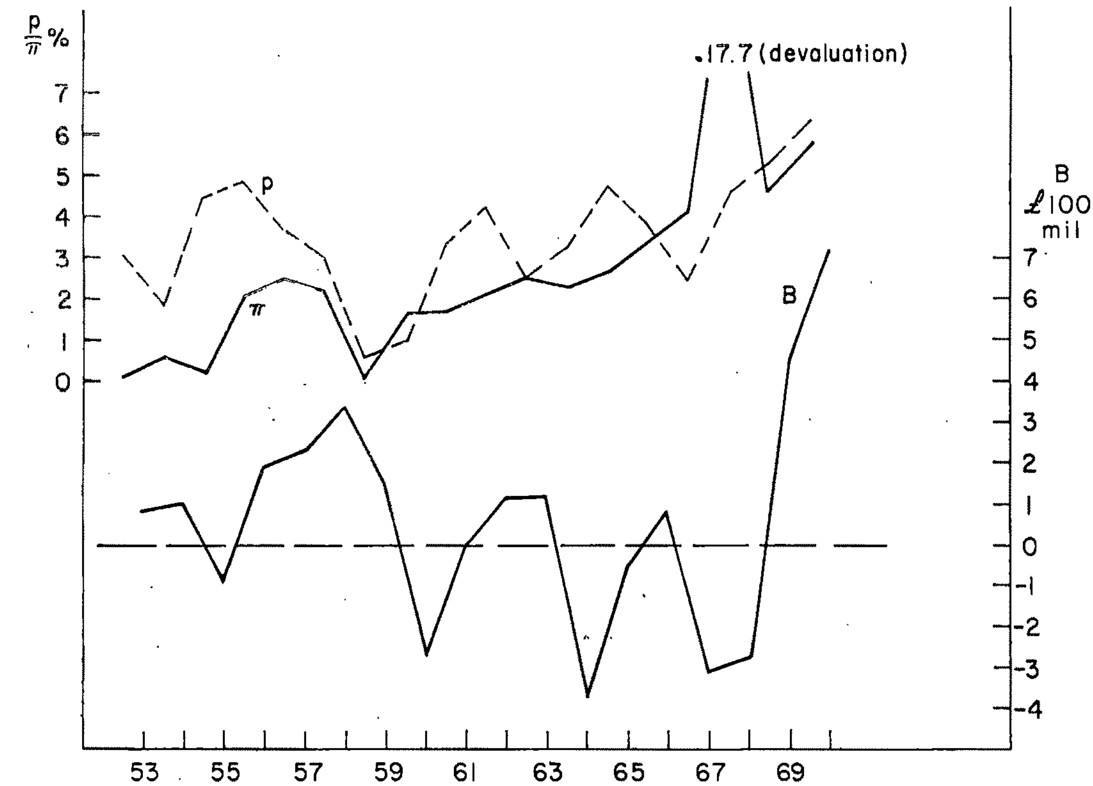



FIGURE 2

$P$ = Annual first differences of the natural logarithm of the retail price index. Basic data are annual averages centered at midyear.

$\pi$ = Annual first differences of natural logarithm of a *GNP* weighted average of various domestic price indices for 19 countries other than the United Kingdom converted at the current exchange rate. Source: Cross and the author where the derivation of the series is described in detail.

$B$ = Balance of payments on current account, including invisibles. *Source: NIESR Review*, various issues.

short-term step up in the inflation rate as British prices adjusted to the new exchange rate, and the lower level of real income implicit in it, even if the inflation rate ruling in the rest of the world had remained constant. But, as Figure 2 also shows, in 1967 the world inflation rate was already accelerating. This acceleration is usually explained as being the result of *U.S.* policies toward financing the Vietnam War.[7] Devaluation and the simultaneous step up in the world inflation rate largely account for the acceleration of British in-

[7] British economists do not always attribute the worldwide inflation of the late 1960's and 1970's to this cause. Thus Sir Henry Phelps-Brown, while admitting that monetary expansion might have been a permissive factor, puts the acceleration of inflation on a worldwide scale down to wage push factors based upon rising real income expectations. Such expectations are transmitted across national boundaries by the motor car and television, these playing the same role as the bicycle, the popular press, and the cinema in the period 1899–1914. He views the Paris students' revolt of 1968 as a crucial event that "sparked a prairie fire of strike action that spread across Europe" (pp. 5–6). Lord Kahn, on the other hand, believes that the breakdown of wage and price controls in Holland in the early 1960's was the source of accelerating worldwide inflation. As he told the House of Commons Expenditure Committee, "In fact it was really from the Netherlands in the second half of the 60s that the infection of wage inflation started and spread to most of the main industrial countries" (p. 82 in House of Commons, 1974, para. 233).

flation after 1967.

Figure 2 shows that the balance of payments did not respond immediately to the 1967 devaluation. Whether this was due to slowness of quantities of imports and exports to respond to relative price changes or to the fact that British monetary policy continued to be expansionary after devaluation is not of central importance.[8] The problem did lead in 1968 to the International Monetary Fund requiring a policy of monetary stringency as a condition of aid, a policy which was adopted and which, it has already been argued, had its effect in 1970–72, at first solely on unemployment and only subsequently on the rate of inflation.

## III

The experience of 1967–71 was widely misinterpreted in Britain, by professional economists and "informed opinion" alike. In the 1960's the Phillips curve in its original form unaugmented by inflationary expectations had gained acceptance among both groups as a valid hypothesis about the relationship between wage inflation and unemployment. For example, at least two of the three operational large-scale forecasting models in terms of which so much discussion of policy took place had, until the late 1960's, treated price inflation as largely depending upon a wage inflation rate determined by some form of Phillips curve. By 1971 the Phillips curve had been dropped from these models leaving money wage inflation as an exogenous variable in all three of them.[9] This change was proximately caused by the failure of the nonexpectations augmented version of the curve to deal with post-1967 data, but also reflected renewed and increasing acceptance of the position, popular in the 1950's, that wage and hence price inflation was the result of wage-push factors of mainly domestic origin.[10] The alternative view that the accelerating inflation of 1967–71 was largely imported with an overlay attributable to the 1967 devaluation found few supporters in 1971.

Crucial to reinforcing the belief that inflation was a matter of domestically originating wage-push was the so-called "wage explosion" of 1969–71. Between 1966 and 1969 the rate of increase of hourly wage rates averaged 5.2 percent per annum; between 1969 and 1971 it averaged 11.1 percent per annum, and that at a time when unemployment averaged about 2.7 percent of the labor force, an unusually high figure by past standards. In retrospect this episode is not difficult to explain. As Table 1a shows, the rate of growth of real wage rates over the 1966–67 to 1970–71 period was similar to rates achieved in the late 1950's and early 1960's. The growth of real wage rates was unusually slow in the mid-1960's. Moreover, as Table 1b shows, price inflation began to accelerate immediately after 1967 and money-wage inflation did not immediately accelerate so rapidly. Some of this wage lag was due to the effect of wage and price controls which had strong short-term effects on wages in 1966–67, and also in 1968 according to some studies (see Parkin and Sumner, 1972, chs. 1, 4, 10). Some of it was perhaps due to lags in the adaptation of expectations to an accelerating and largely imported price

[8] For a detailed analysis of this question, see Peter Jonson and Henryk Kierzkowski.

[9] The three models to which I refer here are those of the Treasury, the National Institute of Economic and Social Research, and the London Business School, respectively. All three treated the wage inflation rate as an exogenous variable, forecast on a judgemental basis, in 1970. (See House of Commons, 1974, para. 456, p. 134; Johnson and Nobay, chs. 1 and 2 with respect to these three models.) It is not clear from the above sources that money wages were ever treated as anything but an exogenous variable in the Treasury model in the 1960's, but Phillips curves definitely were incorporated in the other two.

[10] A typical and comprehensive statement of this point of view in the literature of the 1950's is to be found in Lord (then professor) Kahn's evidence to the Radcliffe Committee. See *Memoranda of Evidence*, Vol. 3, pp. 141–44.

TABLE 1a—AVERAGE ANNUAL PERCENTAGE CHANGE OF PRICES, HOURLY WAGES, REAL HOURLY WAGES, FIVE-YEAR PERIODS, 1956–71

| Period | Prices | Wages | Real Wages |
|---|---|---|---|
| 1956–61 | 2.29 | 4.38 | 2.09 |
| 1957–62 | 2.40 | 4.25 | 1.85 |
| 1958–63 | 2.21 | 4.22 | 2.01 |
| 1959–64 | 2.75 | 4.68 | 1.93 |
| 1960–65 | 3.48 | 5.09 | 1.61 |
| 1961–66 | 3.57 | 5.15 | 1.58 |
| 1962–67 | 3.23 | 5.04 | 1.81 |
| 1963–68 | 3.76 | 5.61 | 1.85 |
| 1964–69 | 4.16 | 5.63 | 1.47 |
| 1965–70 | 4.51 | 6.38 | 1.87 |
| 1966–71 | 5.50 | 7.58 | 2.08 |

TABLE 1b—YEAR ON YEAR CHANGES, 1966–71

| Period | Prices | Wages | Real Wages |
|---|---|---|---|
| 1966–67 | 2.5 | 3.9 | 1.4 |
| 1967–68 | 4.6 | 6.6 | 2.0 |
| 1968–69 | 5.2 | 5.2 | 0.0 |
| 1969–70 | 6.2 | 9.8 | 3.6 |
| 1970–71 | 9.0 | 12.3 | 3.3 |

*Source: NIESR Review*, various issues.

Rate of Change of Money Wages: logarithmic first difference at annual rates of average hourly wage rates in all industries.

Prices: logarithmic first difference at annual rates of retail price index.

Real Wages: col. 2 –col. 1.

inflation. On this interpretation the "wage explosion" simply represents a "catch up" whose underlying causes were strong enough, for a while, to offset the effects of monetary stringency and rising unemployment.[11]

However, this interpretation was very much a minority view at the time. The proceedings of the conference on "The Current Inflation" held at the London School of Economics on February 22, 1971 (reprinted in Johnson and Nobay, 1971) give considerable insight into the state of contemporary opinion. As Harry Johnson put it in his Introduction to the proceedings, "the prevailing mood of the conference seemed to be to dismiss a change in external influences in favor of a change in internal influences, as a primal causal factor in the current inflation" (pp. x–xi). The internal influences in question all centered on the idea of exogenous wage push.[12] By the end of 1971, those in a position to influence British policy were united in the view that the inflation rate was not susceptible to any policy-relevant extent,

[11] Note that an article in the *National Institute Economic Review* (*NIESR Review*) for February 1971, reprinted in Johnson and Nobay (1971) under the name of its author Michael Artis, contained evidence to show that inflationary expectations were systematically influencing wage behavior, though not with a unit coefficient; moreover the relevance of the "Nordic" model of imported inflation (see G. Edgren, K. O. Faxen, and G. E. Ohdner) to the contemporary British situation was considered and recognized. Artis did not, however, develop these ingredients into an analysis of the type offered here.

[12] The following quotations are typical: Jim Ball of the London Business School: "My own favorite candidate is . . . the frustration hypothesis . . . wage earners seek to realize some real wage objectives the frustration of which eventually causes the Phillips curve to shift" (pp. 47–48), though he did note that this hypothesis needed to be squared with the existence of inflation in a number of countries. Leslie Dicks-Mireaux of the Bank of England: "I would not argue that there is an overriding external factor common to the recent inflation; indeed we must probably look to factors closer to home not all of which perhaps are economic" (p. 184). Ralph Turvey, former Deputy Chairman of the Prices and Incomes Board: "it is a question of relative deprivation and the perception of social justice which is the key to understanding what has happened" (p. 200). Sir Fred Catherwood, Director General of the National Economic Development Office: "we are in a position where a key factor of production, which is labour, can be withdrawn and it pays to pay a premium all the time to keep [labor] in full supply. This to me is the major cause of wage inflation" (p. 189). For other recent British statements of hypotheses linking causes of inflation to the institutional structure of the labor market, and other noneconomic factors, see, for example, Aubrey Jones, Sir John Hicks (1974, ch. 3), and Peter Wiles. It is one of the oddities of British debates at the turn of the decade that although the inflationary problems of the time were proclaimed as being something new, requiring equally novel explanations, the explanations that were then actually advanced bore so much resemblance to those which Lord Kahn had put to the Radcliffe Committee more than a decade earlier as explanations of the mild inflation of the 1950's. Moreover, these explanations owe a great deal to the analysis of labor market institutions developed by Hicks in the *Theory of Wages* (1932) and further developed by him in the 1950's. Compare Hicks (1955).

to influence by variations in the level of aggregate demand.[13] Thus, when in that winter, and with a little journalistic license (see Britain, 1975, pp. 70–72), unemployment reached the magic number of 1 million, the scene was set for the implementation of an economic policy which, though coherent enough in terms of this new prevailing orthodoxy, was from a "monetarist" viewpoint inherently contradictory. The strategy, crystallized in Mr. Barber's 1972 budget, involved using fiscal policy to stimulate aggregate demand, and letting monetary expansion take up the resulting public sector deficit lest high interest rates interfere with real expansion or discomfort owner occupiers.[14] Initially an attempt was made to get voluntary controls on wages and prices accepted, but by November 1972 statutory controls were introduced to deal with inflation.[15] Exchange rate flexibility was adopted, explicitly to ensure that balance-of-payments problems did not, as they were judged to have done in the past, jeopardize the pursuit of full employment and growth.

The government were showered with praise for their policy. The National Institute of Economic and Social Research did criticize the budget, it is true, but for not being expansionary enough (see the *NIESR Review*, May 1972). *The Economist*, though noting that the public sector borrowing requirement might lead to monetary expansion, and putting little faith in informal and voluntary wage and price controls, nevertheless concluded that "Within the context of a budget that ignored [getting to grips with wage inflation] Mr. Barber produced something very close to the best economic strategy that a now very professional Chancellor could" (Mar. 25, 1972, p. 13). By November, when statutory controls were introduced, *The Economist* welcomed them with the comment that "It has been obvious for at least the past two years that Britain can avoid South American style inflation only by enforcing a statutory incomes policy" (Nov. 11, 1972, p. 13). No possibility of conflict between rapid expansion of aggregate demand and a policy of wage and

[13] The sharp fall in the inflation rate that was associated with the high unemployment rate ruling in 1971, as the Phillips curve began to reestablish itself, did not have any effect on those who held this view. Indeed it was still the view of the Treasury as late as 1974 that no policy-relevant tradeoff between inflation and unemployment existed. See House of Commons (1974, paras. 465–66, p. 136). In 1974, Patricia Brown of the Treasury explained the fall-off in the inflation rate in 1971–72 in the following terms. "Looking at the rate of growth of the Retail Price Index for the second quarter 1970 to the second quarter 1971, and then the second quarter 1971 to the second quarter 1972, there is a very sharp difference in the rate of growth of food prices . . . the prices of services provided by nationalized industries . . . direct to consumers . . . went up at 17½ percent over the first twelve-month period and only 7½ percent in the second twelve-month period. I think that these two factors probably largely account for the slowing down in the Retail Price Index as a whole" (p. 137 in House of Commons, 1974, para. 69). Perhaps the reader will agree that there is more than a little confusion here between describing what happened to the components of the index on the one hand, and explaining the behavior of its overall value on the other.

[14] 1971 did of course see the advent of new regulations for the Banking System that at least paid lip service to enabling the authorities to exert greater control over monetary aggregates. Whatever the purpose of "Competition and Credit Control," from the very outset the authorities made it clear that steps would always be taken to protect the level of mortgage interest rates under the new regime. See Bank of England *Quarterly Bulletin*, June 1971, para. 15, p. 192. The advocates of variable interest mortgages in the United States should learn from British experience, where such debts are typical, that their existence ensures that all mortgaged homeowners, and not just new borrowers, acquire a vested interest in the maintenance of low interest rates thus creating an immensely powerful pro-inflationary political pressure group. I believe that the desire of all parties to placate owner-occupiers (over 50 percent of all households) played an important role in the political processes that underlay Britain's failure to get to grips with her inflationary problems in the early 1970's.

[15] Note that by 1972 there existed a large body of evidence that suggested that with the exception of the 1948–49 episode, wage and price controls had had no systematic influence on inflation. See Parkin and Sumner, ch. 1, for a survey of this evidence, and note particularly that by then, the initial finding of Richard Lipsey and Parkin (ch. 4 of Parkin and Sumner) that controls did reduce the rate of inflation in conditions of high demand had been discredited.

price controls was noted, not just because inflation was regarded as independent of the level of aggregate demand in the economy, but because expansion of real output and consumption were actually regarded as necessary prerequisites for the success of controls.

An increase in output could be relied on to reduce unit labor costs and hence lead to a short-term slackening off of price inflation; in the longer term an expansion of real consumption would enable the living standards of the labor force to be raised, hence fulfilling the very expectations of rising real incomes the frustration of which, in conditions of slow growth, was supposed to be the main cause of inflation. As the *NIESR Review* for August 1972 put it, "The present situation in the British economy permits a rapid expansion of output for the next two or three years and requires, in the short period at any rate, a sharp increase in consumption. It is the most favorable possible situation for the introduction of [incomes] policies . . ." (p. 6). In November, it returned to the same theme: "The prospects for success of Phase II [of the newly introduced policy] are themselves dependent on the growth of the economy and *on our forecast* some additional stimulus is likely to be needed if the Chancellor's target [of 5 percent real growth] is to be achieved. There certainly seems little need for actually *cutting* public expenditure plans for 1973" (p. 6).[16]

Mr. Barber's "go for growth" policy of 1972–73 was qualitatively similar to Mr. Maudling's 1963–64 "dash for growth," and based on similar intellectual foundations. The Maudling policy quickly foundered on balance-of-payments problems. In 1972, the adoption of exchange rate flexibility against the background of the crumbling Bretton Woods system ensured that the consequence of Mr. Barber's policy would instead be a falling exchange rate and accelerating domestic inflation. However, the adoption of exchange rate flexibility was, in particular, singled out for praise, a typical reaction being that of *The Economist*, which referred to "an extraordinary conversion [of the Government] to the sensible doctrine that the right way to meet any balance of payments deficit will be to set sterling floating" (Mar. 25, 1972, p. 11).

Given only the existence of a stable aggregate demand for money function in the United Kingdom, and there was abundant evidence of this by 1972, there was no conceivable way in which the Conservative Government's policy could have succeeded in achieving the goals of price stability and rapid growth at a high level of employment.[17] In 1972 the money supply was al-

[16] The recipe of high employment and rapid growth as a cure for inflation has, of course, been a long-standing theme in postwar British economics. Thomas (now Lord) Balogh's evidence to the Radcliffe Committee dealt with it in some detail. See *Memoranda*, Vol. 3, pp. 40–41. In fairness to Balogh, however, it should be noted that in 1958 his call was for an expansion of *investment* to stimulate long-term *growth* as a prerequisite for a *voluntary* incomes policy. The 1972 budget concentrated almost entirely upon generating an expansion of *consumption* while the incomes policy that eventually went with it was *statutory*. There was more than a little confusion between long-term growth of productive capacity on the one hand, and short-term expansion up to a level of output that fully utilized existing capacity on the other, underlying economic policy in 1972–73. More than a vestige of the views under discussion here is to be found in a recent paper by Lord Kahn who argues that low unemployment reduces inflationary pressures by making the Trade Unions more willing to cooperate with wage and price controls.

[17] Quite apart from the early studies of Arthur Brown and A. M. Khusro, and Paish's evidence to the Radcliffe Committee, ignored by them in their *Report* (*Memoranda*, Vol. 3, pp. 182–88, *Minutes of Evidence*, pp. 693–700), by early 1972 there were in print studies by Noel Kavanagh and Walters, Douglas Fisher, the Bank of England (1970a), and the author and Parkin, all of which confirmed the existence of a stable demand for money function. There also existed at least three studies that seemed to show that monetary policy variations were associated (albeit weakly) with fluctuations in nominal income with a long and variable time lag (Artis and Nobay, Bank of England, 1970a, b). None of these studies had any influence on policy apparently because the relationships which they produced were not sufficiently well determined to be useful in short-term forecasting exercises. For the Treasury's view on this, see House of Commons (1974, paras. 477–84, pp. 138–39). Recent work by Artis and Mervyn Lewis suggests

ready expanding rapidly, largely as a result of a favorable balance of payments, but also as a result of the implementation of the Competition and Credit Control reforms which had left the banking system with substantial excess reserves. The large public sector borrowing requirement (about 4 percent of *GDP*) implicit in the 1972 budget, coupled with a commitment to hold down interest rates, implied a strong, domestically originating, stimulus to monetary expansion. Under a fixed exchange rate this would have been bound to wipe out the existing surplus and run the balance of payments into deficit, exactly as had happened in 1964, but in 1972 the commitment was to let the exchange rate go in such circumstances.

The possible outcome of the Conservative Government's policy lay between two-not-very-distant extremes. On the one hand, wage and price controls would be ineffective from the outset, domestically originating monetary expansion would lead first to an intensification of an already incipient expansion of real output, and thence to inflation and a falling exchange rate. On the other hand, controls might have some initial success. In this case monetary expansion, not being directly absorbed by rising domestic prices, would spill over into the foreign sector driving the exchange rate even further down than it would fall if controls were ineffective. The result in this case would be an increasingly distorted relative price structure in the economy which eventually would force the abandonment of controls. In either case the inflation rate was bound to accelerate and sterling bound to depreciate. The inevitable duly began to happen in 1972. The monetary expansion rate continued to rise; unemployment started to fall rapidly; inflation to accelerate, slowly at first; the balance of payments went into deficit; and the exchange rate began to fall, reaching a level 10 percent below its trade weighted average Smithsonian parity with other currencies by the end of 1972 and 18 percent below by the end of 1973.

## IV

The conventional and widely accepted interpretation of the 1972–73 boom in Britain is that a correctly conceived policy, reasonably well implemented, foundered as a result of imported inflationary pressures beyond the control of the British authorities. As *The Economist* explained the matter at the time: "Rising import prices and devaluation robbed Mr. Heath of success in stages 1 and 2 [of his incomes policy] and threaten his growth strategy through the impact of high interest rates and monetary restraint to defend sterling" (Oct. 6, 1973, p. 87). *The Economist* did not consider the possibility that "devaluation" might have been the consequence of the government's "strategy" and nor do any of the other proponents of the foregoing interpretation of the events of 1972–73.

Cost accounting exercises do indeed show that over this period rising import prices "contributed" more to domestic inflation than did changes in money wages, profits, or indirect taxes. Such exercises, interesting though they are for the evidence they generate about the behavior of relative prices, are in Britain often interpreted as saying something about the proximate causes of inflation. A clear statement of the approach to the analysis of inflation that leads to such an interpretation was given by G. D. N. Worswick, the director of the National Institute, to the House of Commons Public Expenditure subcommittee: "The principal [factors determining the movement of prices] are the level of wages and costs. The secondary one is profit margins. . . . Thirdly there is the intervention by government by taxes or subsidies to raise or lower the market price from the factor cost price. Fourthly

that competition and credit control did not render the demand for money function unstable after 1971.

there is the import cost. In the last two or three years there has been a world wide increase in the rate of inflation. . . . But who is causing it and how we interact with each other is complicated. We suffer from the world wide rise in commodity prices. . . . The position we started with with the freeing of the exchange rate allowed it to fall rather in comparison with our competitors, which adds to the import prices. To blame that on the incomes policy introduced in 1972 would be, in a certain sense perverse" (House of Commons, 1974, para. 115, p. 38). The Treasury's evidence to the same committee (paras. 462–69, pp. 135–37) shows that their approach to the analysis of inflation is the same as that of Mr. Worswick and *The Economist*, though they did not apply the analysis specifically to the 1972–73 period.[18]

It is above all the behavior of the unemployment rate during the 1972–73 boom that has led so many to believe that inflation since 1972 could not have its source in domestic monetary and fiscal policies and hence must be "imported." As Figure 1 shows, inflation began to accelerate when the level of unemployment was high by historical standards, and even at the peak of the boom it stood above 2 percent, a rate higher than the *average* for the 1953–67 period. This is treated as prima facie evidence that there existed considerable spare capacity in the economy throughout 1972–73.[19] Nevertheless, it is not difficult to reconcile the existence of a "high" unemployment rate with a demand induced explanation of accelerating inflation.

First, and least important, the mechanics of an expectations-augmented Phillips curve, when expectations are based on some kind of error learning mechanism, predict that when unemployment is falling, inflation will accelerate before the natural unemployment rate is reached. There is evidence that such an expectations scheme fits British data rather well.[20] Second, if as many believe, wages and prices respond more rapidly to excess demand than to excess supply, then an expansion of demand that is accompanied by an increase in the dispersion of demand pressure across micro markets will lead to inflation rising above its expected rate *before* the natural unemployment rate is

[18] In its November 17, 1973 issue *The Economist* produced data purporting to show that domestically generated inflation in Britain over the previous year had been lower than in any other OECD country, attributed this to the success of wage and price controls, and headlined the article "The miracle that Ted [Heath] pulled off, and nobody saw him do it." On July 27, 1974, *The Economist* again argued that "Mr. Heath's success in holding [wages] back with his counter-inflation policy begun in the autumn of 1972 is striking. But as wage increases slowed down, import prices took off" (p. 85). For an example of a similar cost accounting exercise masquerading as an explanation of inflation in a learned journal, see R. G. D. Allen. Hicks takes a position somewhat similar to Worswick. He argues that Britain's problem in the 1970's is imported inflation, that the rises in commodity prices that took place in world markets in the 1970's were largely independent of the inflationary policies pursued in industrialized countries, hence were an exogenous source of inflation in those countries, and were not to be counteracted with monetary policy. Even so, his own calculations show that the British import price index which showed an increase from 100 to 1970 to 226 by the end of 1974, would, if recalculated in terms of Deutschmarks, have reached a value of 148 over the same period. In short, 78 percentage points of the change in the index may be attributed to the behavior of exchange rates. See Hicks (1975, p. 10).

[19] The National Institute of Economic and Social Research devoted pages 24–33 of its November 1973 *Review* to the question of whether or not the economy was "overheated," and concluded that there was no reason to suppose that it was. Though not stated in such terms the conclusion was that the natural unemployment rate had not shifted since the 1960's and that the economy was operating well above that natural rate in mid-1973. There is no space here to enter into a critique of the means whereby the *NIESR* reached this conclusion; suffice it to say that I did not then, nor do now, find the relevant arguments convincing, but certainly recognize that this matter is crucial to the interpretation of the 1972–73 boom that I am advancing here. Note that Gray, Parkin, and Sumner estimate that the natural unemployment rate had risen from under 2 percent in the 1950's and 1960's to over 3 percent of the labor force by 1973, but also note that this estimate is subject to a wide margin of error.

[20] See equations (1)–(3) above, although that formulation of the curve is more relevant to a fixed exchange rate regime. Note also that John Carlson and Parkin show that a species of adaptive expectations explains survey data on British inflationary expectations rather well.

reached. There is evidence that the 1972–73 boom was accompanied by just such an increase in dispersion of demand both between industries and regions.[21] Finally, the natural unemployment rate of the British economy was by 1972 well above the level that had prevailed in the 1950's and 1960's. As Figure 1 shows, unemployment rose secularly throughout the period 1953–75, but this tendency was particularly marked after about 1966.

This increase in the natural unemployment rate was widely discussed in Britain in the late 1960's and early 1970's, the phenomenon usually being referred to as a shift in the vacancy-unemployment relationship. It seems to have had a number of sources. The low birth rate of the interwar years coupled with the high postwar birth rate meant that by the late 1960's, older workers reaching retirement age and young relatively inexperienced workers made up an unusually high proportion of the labor force. Moreover, there is some evidence to suggest that until the late 1960's labor hoarding by firms involved the existence of a certain amount of hidden unemployment. This unemployment became visible not only as a result of a "shake out" of labor during the rather deep 1970–71 recession but also as a result of a considerable improvement in the economic status of the unemployed that was brought about in the mid-1960's. Further, there is evidence to suggest that the latter factor contributed to an increase in search unemployment.[22]

Nevertheless the behavior of unemployment over the 1972–73 period simply confirmed those who held it in their belief that inflation was not the result of domestic policies. The *NIESR* went so far as to interpret the behavior of the balance of payments, which had gone from a current account surplus of £0.07 billion in 1972 to a deficit of £1.47 billion, or over 2 percent of *GDP*, in 1973 despite an 18 percent fall in the exchange rate over the two previous years, as providing evidence *against* the presence of general excess demand in the economy.[23] However, even if the deterioration of the balance of payments and the decline in the exchange rate were not attributed to the conduct of domestic policy by the authorities and their advisors, these factors certainly began to influence the conduct of policy at some time in 1973. Although the miners' strike, the three-day week, and the February election marked the final collapse of the Conservative Government's policy in early 1974, that policy had been changing for a few months previously. Although expansionary policies were not blamed for the behavior of the balance of payments and the exchange rate, contractionary policies were nevertheless implemented in order to deal with these problems.

## V

We have seen that monetary policy became sharply contractionary in 1973, and continued in that vein throughout 1974. This change roughly coincides with a large increase in the share of public sector borrowing financed by sales of debt to the nonbank public. This ratio averaged 35 percent over the last three quarters of 1972 and 55 percent in 1973. Whether the reversal of the monetary aspect of economic policy was deliberate or accidental is hard to say, for it has been and remains the constant claim of all political parties that they

[21] See the *NIESR Review*, Nov. 1973, pp. 28–29, for a discussion of the dispersion of aggregate demand in 1973.

[22] On demographic factors, see Foster; on labor hoarding see Taylor; and on the role of the level of unemployment benefits see Dennis Maki and Z. A. Spindler. Samuel Brittan provides a useful summary of the evidence on these matters.

[23] See the *NIESR Review*, Nov. 1973, p. 30. The case put there rests on the volume of imports and exports having grown at the same pace during the boom, whereas it is argued that had there existed domestic excess demand, the volume of imports should have increased more than that of exports. The *NIESR* did not consider the behavior of the exchange rate relevant to interpretation of this evidence, and did not explain why.

would not resort to a tight money policy in order to combat inflation, but this turnaround was shortly followed by an attempt to reverse fiscal policy as well. The borrowing requirement of the public sector for 1973–74 generated by the March 1973 budget was just under £4½ billion or about 7 percent of *GDP*. A supplementary budget introduced in December 1973 sought substantially to reduce the borrowing requirement for 1974–75, and at the beginning of March 1974 it was forecast to be £3.4 billion. Although some observers were worried that the reversal was being implemented a bit too rapidly for comfort, particularly as far as the rate of monetary expansion was concerned, it nevertheless appeared that in the winter of 1973–74, some kind of rough and ready sanity was being restored to the conduct of macroeconomic policy in Britain despite prevailing economic orthodoxy. In fact, the very opposite was happening.

Budgets are usually an annual affair in Britain, but the December 1973 budget was the first of a series of five in a sixteen-month period. Even the most orthodox Keynesian might find such zeal for fine tuning excessive, and suspect that all was not well with the conduct of policy during these months. However, the actual aggregate outcome of these budgets bore so little resemblance to the government's expressed intentions at the time of their introduction that the only reasonable conclusion can be that 1974 saw the disintegration of the machinery of macro-economic policymaking in Britain. During the fiscal year 1974–75, the public sector intended to borrow just under 5 percent of *GDP* but in fact ended up borrowing about 10 percent; public sector borrowing for 1975–76 is running at about 12 percent of *GDP* and over one-quarter of this borrowing again was not forecast. This unintentionally highly expansionary fiscal policy ran against a continued tight monetary policy, and the unemployment statistics for 1974–75 are eloquent testimony as to which policy tool proved the more powerful.

At this stage there can be no definitive account of what went wrong with the budgetary process in 1974–75, but two broad factors seem to have contributed to the debacle, the first political and the second technical. The Labor government elected in early 1974 replaced statutory wage controls with a voluntary program known first as the "Social Compact" and later as the "Social Contract." Its essence was that the trade unions agreed to limit their demand for increased wages to the rate of inflation, while the government in turn agreed to maintain for a while the price control program of their predecessors, to freeze certain key prices such as rents, and to subsidize certain basic foodstuffs. Since it was never specified whether the rate of inflation to which wage demands were to be limited was the past rate, the current rate, or the expected rate, it was not even necessary for anyone to resort to subterfuge in order to achieve whatever wage increases market conditions indicated.

General wage inflation coupled with price restraint in the public sector involved nationalized industries in increasing deficits; food subsidies had to be paid; and a rent freeze in an economy where about one-third of households occupy already subsidized publicly owned housing ensured a further open-ended commitment to public expenditure. One cannot lay the social contract at the door of professional economists, either inside or outside of government service.[24] They are more culpable of

[24] However in August 1974 the *NIESR*, though doubting that the Social Contract would achieve all its aims, was nevertheless arguing that it be given a year to run as a prelude however to a new round of statutory controls (p. 6). In House of Commons (1974), responding to a question about the desirability of statutory wage and price controls, Sir Kenneth Berrill, then Chief Economic Advisor at the Treasury, is recorded as having said, with more than a little irony, "I do not think I am able to answer that because the present view of the Government, and, therefore, of the Treasury, is that a voluntary incomes policy will work" (p. 141 in House of Commons, 1975, para. 496).

whatever it was that went wrong with official and unofficial forecasts of inflation and unemployment rates in 1974–75. Public expenditure planning in Britain is carried out in real terms. Government departments are automatically granted funds to cover inflation induced cost increases. There is no current purchasing power cash budget to constrain their expenditure. Such a *modus operandi* means that errors in forecasting the inflation rate and relative prices play havoc with government finances. At the same time, the whole thrust of British government forecasting is geared to predicting real income and unemployment. Forecasts of the inflation rate are ancillary to this goal, so that it is not too surprising that such forecasts are relatively unreliable. The major part of the error in forecasting public sector borrowing in 1974–75 seems to have arisen from underestimates of the effects of inflation on public sector costs.[25]

Even though public sector borrowing climbed steadily in 1974–75, and is doing the same in 1975–76, a relatively low rate of monetary expansion was nevertheless maintained until the end of 1975. Although the current account of the balance of payments remained in heavy though decreasing deficit throughout the period, the exchange rate nevertheless remained stable between the end of 1973 until mid-1975 when it fell abruptly to an effective devaluation of about 30 percent below its Smithsonian parity. The explanation of these facts is straightforward. First, the ratio of sales of public sector debt to the nonbank public to total public sector borrowing remained high throughout 1974 and 1975. Moreover, the oil price increase of 1973 led to large capital inflows to Britain, first as oil producers invested surplus funds in London, and secondly as the increase in the prospective value of North Sea oil attracted development capital. In addition to this the public sector actively raised funds in foreign markets over this period.

The effects of the exchange rate being thus supported by capital inflows and public sector borrowing were to permit real domestic expenditure to remain substantially above domestic output and to ensure that the inflation rate was somewhat lower than it otherwise would have been. Policies to reduce the level of public sector borrowing must form a key part of any strategy that is to restore long-run stability to the British economy, and the period of exchange rate stability during 1974–75 might have played a valuable role in smoothing out the inflation rate a little while such policies were implemented. However, this breathing space was wasted; all that was accomplished by the delay in the decline of the exchange rate was to transfer to late 1975 and early 1976 some inflationary pressure that would otherwise have been felt earlier. In 1974–75, rising unemployment generated by, but not attributed to, monetary policy inhibited the authorities from any serious attempts to tighten up on the fiscal side, though given the breakdown in the machinery of policy implementation already described, it is far from clear that they could have succeeded in this endeavor if they had tried.

## VI

The British economy at the beginning of 1976 presents a confusing picture. The public sector is borrowing over 12 percent of *GDP*, much of this to finance transfers and subsidies. Increasingly this borrowing requirement must be met domestically. At the same time, unemployment, though rather low by current international standards, is still high enough to create a major political problem. The government is un-

[25] See House of Commons (1975) for an account of these and other factors affecting public sector borrowing. I base my account of the role of the inflation rate in official forecasting on the Treasury's evidence to the House of Commons Public Expenditure Committee. See House of Commons (1974, paras. 426–33, pp. 127–30 and paras. 455–63, pp. 134–35).

dertaking all manner of piecemeal job saving and creating schemes to deal with it and these are, of course, adding further to public expenditure. Inflation still seems to be regarded in official circles as mainly a cost-push phenomenon, and recently introduced wage controls are going for a while to seem successful as unemployment, generated by the tight monetary policy set in motion in 1973, has its predictable effect on the inflation rate. The one hopeful aspect of the current situation is that far more attention is now being paid to the behavior of the money supply than in the past, and the government has recently and publicly committed itself not to permit renewed rapid monetary expansion. The question remains, however, as to how durable that commitment will prove to be.

The key to maintaining a reasonable monetary expansion rate in Britain must be a reduction in public sector borrowing. It is inconceivable that this can continue to run at the rate of 12 percent of *GDP* without thereby generating a renewed acceleration in the rate of money creation.[26] The inflation rate is now falling rapidly, and an upturn in real economic activity is forecast for 1976. Both of these factors should help to bring the borrowing requirement down, and there seem to be no insuperable technical problems involved in preventing a large increase in the monetary expansion rate.

The real danger in the present situation is political. There is no sign of any early fall in the unemployment rate, and the government must find itself under increasing pressure to "do something" about this problem. If those in authority are convinced that inflation has been brought down in 1975–76 not by monetary policy, but rather by the use of wage controls, and if they are also persuaded that high unemployment makes such controls less rather than more effective, as is still being argued in some circles,[27] then they will be strongly tempted to implement expansionary fiscal policy, accompanied by wage and price controls, whose consequences for monetary expansion and the inflation rate will be qualitatively similar to those of the 1972–73 episode, even if resort is had to import controls in an attempt to shore up the exchange rate and hence to offset "imported" inflation. In short, whether 1975 has seen the peak in Britain's postwar inflation, or merely an upper turning point in a continuing and explosive series of cycles remains to be seen.

[26] It is fair to record that this seems to be the view of *The Economist*, which now pays much more attention to the role of monetary factors in generating inflation than it did even as late as 1974. See, e.g., January 24, 1976, where it is noted that "Mr. Denis Healey will keep tighter screws on money and demand than any of his postwar predecessors during [1976]. So far so better" (p. 71).

[27] See, for example, Lord Kahn (1976).

## REFERENCES

**R. G. D. Allen,** "The Immediate Contributors to Inflation," *Econ. J.*, Sept. 1975, *85*, 607–11.

**M. J. Artis and M. K. Lewis,** "The Demand for Money in the U.K. 1963–73," *Manchester Sch. Econ. Soc. Stud.*, forthcoming.

**M. J. Artis and A. R. Nobay,** "Two Aspects of the Monetary Debate," *Nat. Instit. Econ. Rev.*, Aug. 1969, *49*, 33–42.

**S. Brittan,** *Second Thoughts on Full Employment Policy*, London 1975.

**A. J. Brown,** "Interest, Prices and the Demand for Idle Money," *Oxford Econ. Pap.*, May 1939, *2*, 46–69.

**J. A. Carlson and J. M. Parkin,** "Inflation Expectations," *Economica*, Feb. 1975, *42*, 123–38.

**R. J. Cross and D. Laidler,** "Inflation, Excess Demand and Expectations in Fixed Exchange Rate Open Economies: Some Preliminary Empirical Results," in J. M. Parkin and G. Zis, eds., *Inflation in the World Economy*, Manchester 1976.

**G. Edgren, K. O. Faxen, and G. E. Odhner,** "Wages, Growth and the Distribution of Income," *Swedish J. Econ.*, Sept. 1969, *71*, 133–60.

**D. Fisher**, "The Demand for Money in Britain: Quarterly Results 1951–67," *Manchester Sch. Econ. Soc. Stud.*, Dec. 1968, *38*, 327–51.

**J. I. Foster**, "The Relationship Between Unemployment and Vacancies in Great Britain 1958–72: Some Further Evidence," in D. Laidler and D. Purdy, eds., *Inflation and Labour Markets*, Manchester 1974.

**M. Gray, J. M. Parkin, and M. T. Sumner**, "Inflation in the United Kingdom; Causes and Transmission Mechanisms," mimeo., Univ. Manchester 1975.

**J. R. Hicks**, *The Theory of Wages*, London 1932.

———, "Economic Foundations of Wages Policy," *Econ. J.*, Sept. 1955, *65*, 389–404.

———, *The Crisis in Keynesian Economics*, Oxford 1974.

———, "What is Wrong with Monetarism," *Lloyds Bank Rev.*, Oct. 1975, no. 118, 1–13.

**H. G. Johnson**, *Inflation and the Monetarist Controversy*, Amsterdam 1972.

——— **and A. R. Nobay**, *The Current Inflation*, London 1971.

**A. Jones**, *The New Inflation. The Politics of Prices and Incomes*, London 1972.

**P. D. Jonson and H. I. Kierzkowski**, "The Balance of Payments: An Analytic Exercise," *Manchester Sch. Econ. Soc. Stud.*, June 1975, *43*, 105–33.

**R. F. Kahn**, "Inflation, a Keynesian View," *Scottish J. Polit. Econ.*, Feb. 1976, *23*, 11–16.

**N. J. Kavanagh and A. A. Walters**, "The Demand for Money in the United Kingdom 1877–1961: Preliminary Findings," *Bull. Oxford Univ. Inst. Econ. Statist.*, May 1966, *28*, 93–116.

**A. M. Khusro**, "An Investigation of Liquidity Preference," *Yorkshire Bull. Econ. Soc. Res.*, Jan. 1952, *4*, 1–20.

**D. Laidler**, *Essays on Money and Inflation*, Manchester; Chicago 1975.

———, "Alternative Explanations of and Policies towards Inflation in Six Countries," in K. Brunner and A. H. Meltzer, eds., *Proceedings of March 1975 Carnegie-Mellon Rochester Conference on Economic Policy*, forthcoming.

——— **and J. M. Parkin**, "The Demand for Money in the United Kingdom 1955–1967 Preliminary Estimates," *Manchester Sch. Econ. Soc. Stud.*, Sept. 1970, *38*, 187–208.

**D. R. Maki and Z. A. Spindler**, "The Effect of Unemployment Compensation on the Rate of Unemployment in Great Britain," disc. pap. 7413, mimeo., Simon Fraser Univ. 1974.

**J. M. Parkin and M. Sumner**, *Incomes Policy and Inflation*, Manchester 1972.

**E. H. Phelps-Brown**, "A Non-Monetarist View of the Pay Explosion," *Three Banks Rev.*, June 1975, no. 105, 183–91.

**A. W. Phillips**, "The Relation Between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom," *Economica*, Nov. 1958, *25*, 283–99.

**J. Taylor**, "Incomes Policy, the Structure of Unemployment and the Phillips Curve: The United Kingdom Experience 1953–70," in J. M. Parkin and M. T. Sumner, eds., *Incomes Policy and Inflation*, Manchester 1972.

**A. A. Walters**, "Money in Boom and Slump," Hobart pap. no. 44, London Inst. Economic Affairs, 1969.

**P. J. Wiles**, "Cost Inflation and the State of Economic Theory," *Econ. J.*, June 1973, *83*, 377–98.

**Bank of England**, (1970a) "The Importance of Money," *Bank of England Quart. Bull.*, June 1970, *10*, 159–98.

———, (1970b) "Timing Relationships Between Movements of Monetary and National Income Variables," *Bank of England Quart. Bull.*, Dec. 1970, *10*, 459–68.

———, *Statistical Abstract, 1953-60*, 1971.

**House of Commons**, *Ninth Report from the Expenditure Committee, Session 1974-75, Public Expenditure, Inflation and the Balance of Payments*, London 1974.

———, *First Report from the Expenditure Committee, Session 1975-76, The Financing of Public Expenditure*, Vol. 1, London 1975.

***International Financial Statistics*** (*IFS*), various issues.

**National Institute of Economics and Social Research**, *National Institute Economic Review (NIESR Review)*, various issues.

**Radcliffe Committee** (Committee on the Working of the Monetary System), *Report; Principal Memoranda of Evidence; Minutes of Evidence;* London 1959.

***The Economist***, various issues.

# Can a Rise in Import Prices Be Inflationary and Deflationary?

## Economists and U.K. Inflation, 1973-74

*By* Marcus H. Miller*

In June and July of 1974, the influential Expenditure Committee of the House of Commons heard submissions from economists in the public and private sector on "Public Expenditure, Inflation, and the Balance of Payments," House of Commons (1974). The hearings came just a few months after the sharp rise in oil prices, and the demise of a Conservative government whose incomes policy had made no special allowance for a rise in coal miners' income despite the increase in energy prices. The witnesses called to give evidence included prominent British macroeconomists of both Keynesian and Monetarist persuasions; thus, for example, Lord Kahn and David Laidler both spoke before the Committee.

The Committee reported, "We are told that a rise in the price of imports was both inflationary and deflationary" (para. 23), which they understood to mean "price increasing and employment reducing." They also noted that the various witnesses were far from unanimous in what they thought the impact effects would be, and in what they recommended by way of policy. This paper will initially focus on how Keynesians and Monetarists expected the shift in the terms of trade to affect prices and output, and what policy conclusions were derived, using evidence given to the Committee. The main reason for concentrating on this aspect of recent inflation is evident from inspection of Figure 1 which shows how severe was the shock of the rise in import prices over the period 1973–74 (import unit values rose by over 60 percent from 1973-II to 1974-II).

Later I discuss how the inflationary effects of such an external shock may be amplified by what Sir John Hicks (1975b) has dubbed "real wage resistance," so that a bout of "imported" inflation may be followed by a spell of "home-grown" inflation. Some witnesses (including those from the Treasury and the National Institute) had referred to the possibility of inflationary pressures from this source (House of Commons, paras. 49, 50, 136, 498).

I will argue, however, that neither the Committee nor Hicks gave sufficient attention to the role of the incomes policy in operation when the price of oil rose so dramatically. For this policy not only involved confrontation with the coal miners, it also led to the linking of the wages of about one-third of the work force to the retail price index, at a time when the latter rose sharply because of a change in terms of trade. Thus the incomes policy helped to prevent real wages from falling when economic circumstances called for such a change.

Perhaps it is not surprising, therefore, that the report had very little positive to say about incomes policy, confining itself to expressing the view that "a permanent, statutory prices and incomes policy is in modern Britain politically both imprac-

* London School of Economics and Graduate School of Business, University of Chicago. I would like to thank M. J. Artis, T. Burns, S. G. B. Henry, R. A. Jackman, D. Laidler, D. Sargan, and J. Wise for their comments, without implying that they would accept the views expressed here.

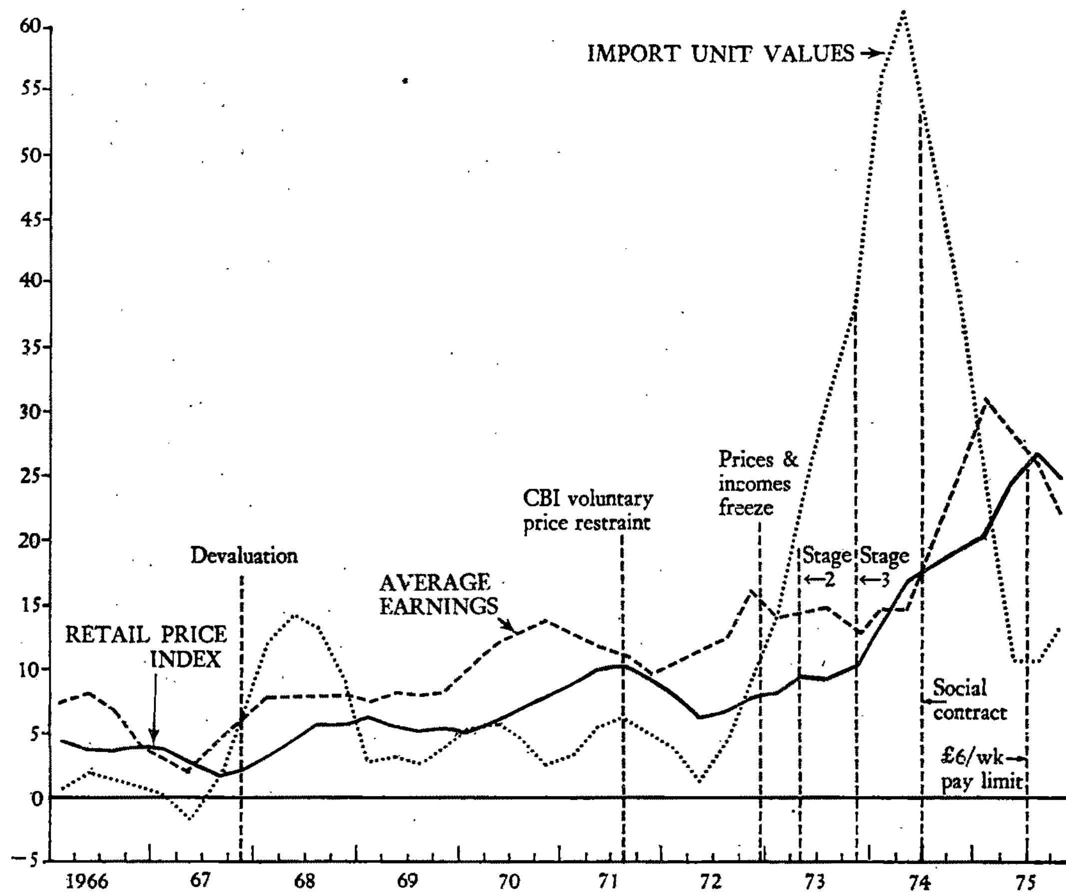



FIGURE 1. RETAIL PRICES, AVERAGE EARNINGS AND IMPORT PRICES (unit values)

*Source: NIESR Review.*

Definitions: Average Earnings (*gB*) (Wages + Salaries Combined) % increase on year earlier
Important Unit Values—% increase on year earlier
Retail Price Index—% increase on year earlier

ticable and objectionable" (para. 44). An incomes policy calling for a dramatic reduction of nominal wage growth was nevertheless implemented by a Labour administration within a year of the report being published, and the future prospects for the exchange rate appear to depend heavily on the appropriate extension of such a policy beyond July 1976. Though I criticize the policy of indexing wages pursued over the period 1973–74, my examination of the determination of real wages indicates a useful role for incomes policies in mitigating the "cost-push" inflationary pressures that arise from increased taxation and from changes in the terms of trade.

## I. The Economic Situation Prior to the Oil Price Rise

Prior to analyzing the consequences of the rise in import prices from mid-1973 to mid-1974, I must briefly consider the situation before then. The economy as a whole was in what has been referred to as a "super-boom" as a consequence of expansionary action on a number of fronts.

As regards fiscal policy, for example,

there had been a substantial increase in the "adjusted budget deficit" since 1971 (see Table 1). The figures show that the adjusted deficit had risen by about $1\frac{1}{2}$ percent since 1971, despite the fiscal drag to be expected during such a recovery. Monetary policy had also been extremely lax since the reform of the monetary system in 1971.[1] Broad money, $M_3$, grew by over 50 percent from mid-1971 to mid-1973, for example.

Soon after it had become apparent that the economy was set for a period of very rapid expansion, there was a run on sterling with the result that the currency was floated in mid-1972. The immediate adjustment was a 10 percent fall in the "effective" rate from its Smithsonian parity, but this was followed by a further depreciation of 7 percent in the first half of 1973. Not surprisingly, in the circumstances described, unemployment fell from $3\frac{1}{2}$ percent in mid-1972 to $2\frac{1}{2}$ percent in mid-1973.

In an effort to contain the undoubted inflationary pressures generated by this combination of fiscal, monetary, and exchange rate stimuli, the Conservative government, much against its inclinations, instituted a three-stage incomes and prices policy as follows:

> A freeze on wages and salaries was introduced in November 1972 (though not until after a large number of double-figure anticipatory settlements had been rushed through in the preceding months) and lasted until the end of March, although pre-freeze agreements, of which there were many, were allowed to stand. This was followed by Stage 2, in which increases of up to £1 per week plus 4 per cent were permitted (with a ceiling of £250 per annum) and then from November 1973 Stage 3, in which the allowable increase was basically £2.25 per week or 7 per cent (with a ceiling of £350 per annum). [*NIESR Review*, 1974, p. 18.]

In addition, and of vital importance for the subsequent course of events, under Stage 3 wages could be "indexed" for twelve months under a threshold agreement which came into operation when the retail price index (*RPI*) rose by more than 7 percent above its level in October 1973.[2]

With import prices rising by 25 percent over the year to mid-1973, earnings rose by 15 percent and prices by 10 percent over the same period, despite the incomes policies (which did allow for certain costs to be passed on). The deflationary budget and the drastic action to control the money supply taken in November 1973 were designed to offset some of the then existing inflationary pressures and are not to be thought of as the response to the oil crisis. The sort of reaction appropriate to a rise in the oil price itself is considered next.

### II. The Keynesian Diagnosis

I turn to the Keynesian account of how the sharp rise in the price of imports at constant exchange rates was "both inflationary and deflationary." G. D. N. Worswick, Director of the National Institute of Economic and Social Research, endorsed this view before the Committee (paras. 132–35) and I will cite extensively from the National Institute's quarterly review in what follows.

I must first warn the reader that a large change in relative prices of this kind plays tricks with key macro-economic measures. Thus, we may have a rise in the domestic

[1] The program of changes was called "Competition and Credit Control." As part of this program ceilings on bank lending were scrapped and what control there was on bank lending was supposed to come from control of "reserve assets"; but these were in plentiful supply for much of the time, and burgeoning government deficits were taken by the banks to indicate that the supply would continue.

[2] The maximum compensation for inflation was at the rate of 40p per week for every 1 percent rise in the *RPI*, which was designed to represent complete compensation for the average male manual wage earner (but was clearly more than that for the lower paid and less for the better paid).

TABLE 1—ADJUSTED PUBLIC SECTOR DEFICIT: 1970–74[a]

| | Weighted Receipts | Weighted Expenditure | Weighted Deficit | Weighted Deficit Less Impact of Oil Deficit[c] |
|---|---|---|---|---|
| 1970 | 25.3 | 33.0 | 7.7 | — |
| 1971 | 24.3 | 33.0 | 8.8 | — |
| 1972 | 23.3 | 33.4 | 10.1 | — |
| 1973 | 23.4 | 33.6 | 10.2 | — |
| 1974[b] | | | | |
| Mar. I | 24.9 | 34.6 | 9.6 | 7.6 |
| Mar. II | 25.7 | 35.0 | 9.3 | 7.3 |
| Nov. I | 25.7 | 37.4 | 11.8 | 9.8 |
| Nov. II | 25.2 | 37.4 | 12.2 | 10.2 |

*Source: NIESR Review* (1975, pp. 14–16).

[a] All figures expressed as percentage of current *GDP*; the weights attempt to measure the effect of budget components on aggregated demand; no attempt is made to "standardize" the budget on full employment income.

[b] I before budget changes; II after budget changes.

[c] The impact of the oil deficit on aggregate demand is put at 2 percent *GDP*.

price level without any change in the *GDP* deflator, and a fall in real incomes without any change in *GDP* at constant prices. This may be seen most clearly by leaving on one side any employment effects, and by assuming also that the rise in import prices leaves domestic wages and profits unchanged in nominal terms. Defining nominal income as $Y \equiv PF - P_I M$, where $F$ is the total final sales, $P$ is the associated price index,[3] $M$ is the volume of imports, and $P_I$ is the price index for imports, we may further assume that $M = mF$, so import volumes depend upon the total final sales. (Ignoring the price elasticity of demand for imports is not unreasonable for the case of raw materials and oil.) Given that sales and import volumes are constant, the domestic price level $P$ may rise as higher import costs are passed on, without any change in domestic nominal incomes; hence the *GDP* deflator is not an appropriate index of inflation in these circumstances, when it is "foreign costs" not "home costs" which have risen.

Nor is constant price *GDP* a good measure of real national income since, with volumes unchanged, such a constant price measure will show no change; yet clearly real incomes fall when prices rise relative to constant nominal income, as can be seen from defining[4] real national income as $Y^R \equiv Y^N/P = F - MP_I/P$. The fall in the nation's real income due to the shift in the terms of trade between 1973 (first half) and 1974 (first half) was estimated by the Central Statistical Office to be over 5 percent (see article by J. Hibbert).

The Keynesian account of the *inflationary* impact of the rise in the price of imports can now be expressed simply enough. Real national income has to fall,

[3] Final sales includes exports (as well as government purchases and private consumption and investment). For the sake of simplicity it is assumed that the price of exports moves in line with that of goods sold domestically, so that one index $P$ can be used for all such sales. The important consequences of the price of exports rising relative to other domestic sales are thus not considered in this paper. The Southampton University econometric model being constructed by I. F. Pearce and his colleagues is specifically designed to examine the effects of such changes.

[4] In answer to questions on this point, Wynne Godley answered as follows:

> The real national output is defined as the total value of what we produce valued at constant prices—that is *including* exports at constant prices but *excluding* imports at constant prices. The real national income is defined as the total factor income deflated by the price of domestic expenditure. [House of Commons, para. 89]

but if nominal incomes are rigid downward at a given level of employment, the fall in real incomes must take place by prices rising relative to factor incomes, as import costs are passed on. Nor would appreciating the exchange rate to prevent any rise in the price of final sales $P$ obviate the need for a fall in nominal factor incomes, since the appropriate upward float would only be a fraction of the initial rise in the price of imports; which is one reason why Keynesians were not inclined to recommend exchange rate changes in this context.

We have seen that (even with floating exchange rates) stability of final output prices would only be achieved by a fall in nominal factor incomes relative to what they would have been in the absence of the shift in the terms of trade. Such a fall would, in the Keynesian view, require a rise in unemployment. Now, as we see below, Keynesians were in fact predicting a rise in unemployment in these circumstances, though their policy prescriptions were to offset this and to accept the "imported" rise in the price level.

On the income-expenditure approach, the fall in real income described above will tend to reduce aggregate demand. As the National Institute reported: "The demand-contractionary impact of the new oil prices (akin to a large increase in indirect taxes except that the proceeds accrue to foreigners) was . . . quickly appreciated in this country" (*NIESR Review*, 1975, p. 9). While the fall in national real income will obviously depress domestic aggregate demand, the net effect of such a "transfer" of income depends crucially on the marginal propensity to consume *U.K.* exports on the part of the recipients of the transfer. While this propensity is high for many raw material producers, it was widely accepted that this would not be so for the Arab oil producers; hence, the net impact of the oil price rise was reckoned to be contractionary.[5] Assuming *no* change in exports, the appropriate expansionary action on the income-expenditure approach is for the government to cut taxes by the fall in national real income, so that real national disposable income remains constant, or to increase its own expenditure by the marginal propensity to spend out of the real income loss (see the Appendix).

This is illustrated in Figure 2 which shows total final sales $F$ on the vertical axis as an increasing function of real income $Y^R$ on the horizontal. The cost of imports, whose volume depends on $F$, is subtracted from $F$ to give a measure of domestic-income-creating expenditure (current $GNP$ deflated by the price of total final sales) on the vertical axis. From a position of equilibrium with the current account balance at full employment (with sales at $F_0$), the terms of trade ($h \equiv P_I/P$) shift from $h_0$ to $h_1$, reducing full employment real national income by $M_0 dh$ from $Y_0^R$ to $Y_1^R$. As we can see, there is a deficiency of aggregate demand at this new level of full employment real income. To avoid the unemployment associated with the equilibrium shown at $Y_1^R$ and total final sales of only $F_1$, the authorities must increase the ratio of expenditure to income, as shown by the schedule $F^*(1-h_1 m)$, through an increase in their own spending or a cut in taxes or some combination of the two.

By splitting total final sales into absorption $A$ and exports $X$, and plotting $Y_R - A$ and $X - hM$ against real income as in the lower panel, it is clear that preserving full employment at $F_0$ in this way involves accepting a deterioration in the cur-

[5] The *NIESR Review* (1975) reported that:

> The new Chancellor was quick to point out to his colleagues in other finance ministries the strength of the contractionary forces involved and urged them—especially those with strong external accounts—to take offsetting reflectionary action . . . for nearly the whole year his exhortations fell on stony ground—the Americans, in particular, delayed in the hope of forcing the oil price down again, while in other countries the implication that acceptance of external deficit necessarily meant the acceptance of greater budgetary deficit . . . was not taken to heart. [pp. 9–10]

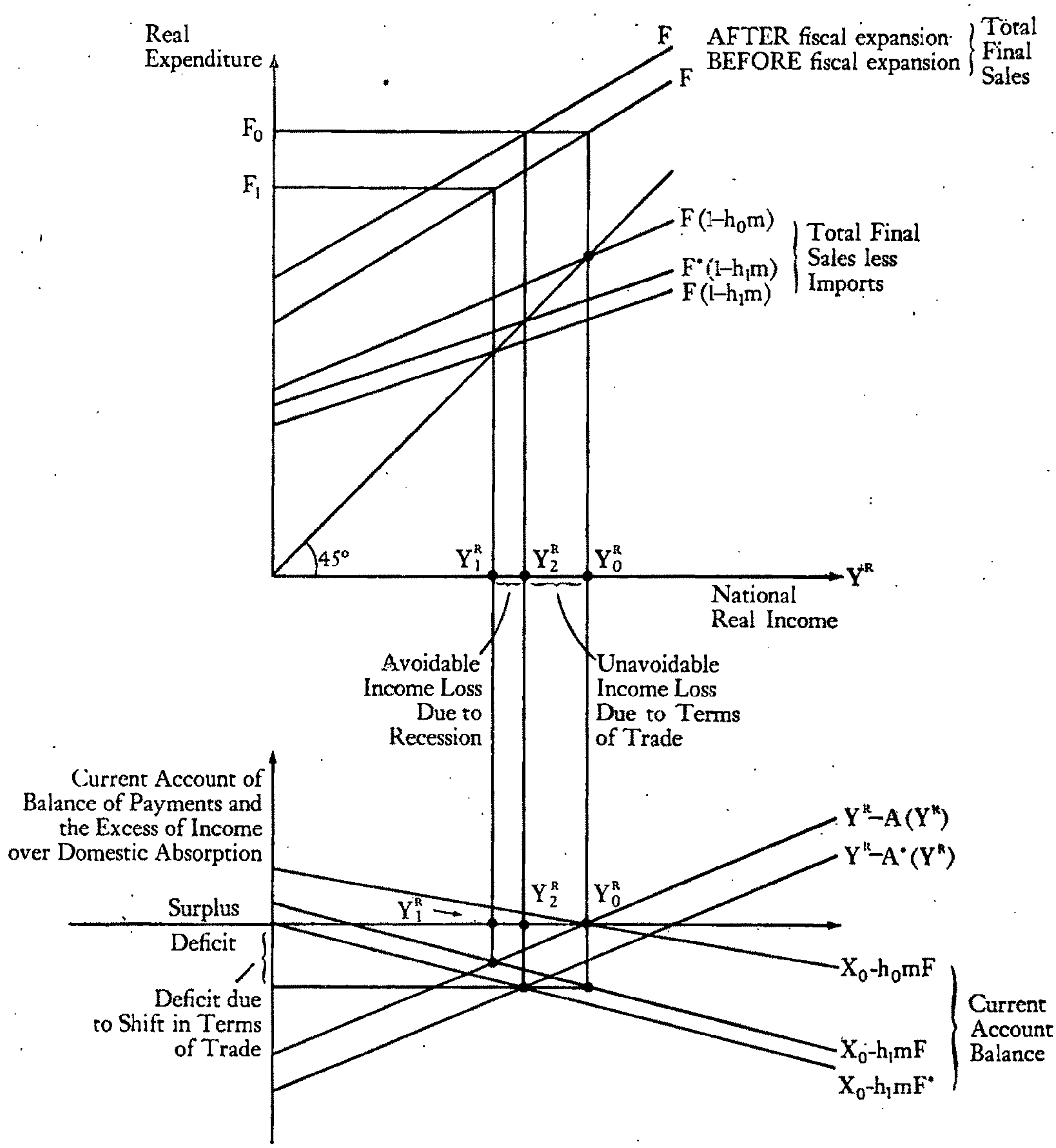



FIGURE 2

rent balance equal to the reduction in real national income.[6]

[6] How much real income will fall in the absence of government action obviously depends on the slopes of the two schedules. One of the principal issues investigated by the committee was the contention by Wynne Godley and Francis Cripps of Cambridge University that the schedule showing the excess of real income over absorption would not increase markedly with real income, implying a large multiplier effect of import price changes on domestic income and employment. The committee concluded only that Godley's hypothesis "was not obviously discredited" (House of Commons, para. 37).

As for the fiscal action actually taken by the *U.K.* authorities in 1974, the *NIESR Review* (1975) noted that:

The Chancellor's intention actually to

> *reduce* the U.K.'s budget deficit in his (admittedly provisional) March budget appeared to run counter to his correct analysis of the situation of the industrial world as a whole. . . . The subsequent budgets of July and November and, most importantly, the revision of the March estimates, restored consistency between the Chancellor's words and his deeds. [p. 10]

As can be seen from Table 1, the increase in the adjusted budget deficit between 1973 and the end of 1974 just matched the deflationary impact of the oil deficit on *U.K.* demand (put at 2 percent by the National Institute). Although the increase in the weighted deficit between March and the end of the year was to a large extent "nondiscretionary," the fact that the authorities did not act to reverse this increase in the November budget shows that they were willing to act in such a way as to offset the contractionary impact of the oil price increase.[7]

### III. The Monetarist Diagnosis

While Keynesians were in favor of letting the rise in prices reduce real incomes and were more exercised about the level of employment, Monetarists scoffed at the idea that there was any need to "import inflation" in this way and recommended a policy of not increasing "monetary demand" in response to the adverse change in the terms of trade. But if final prices are not to rise when, as a result of the change in the terms of trade, real incomes have fallen, then, as we have seen, nominal incomes must fall below the path that they would otherwise have followed. This is not to say that Monetarists expected money wages to fall; simply that their principal policy recommendations, if they were not to lead to unemployment, required that money wages would rise *less*—in the presence of a shift in the terms of trade—than they would otherwise have done.

Monetarists were, moreover, not particularly concerned about the effect of such a slowdown in incomes on aggregate demand, since of course for them the principal determinant of nominal expenditure is the money supply, not nominal income. In two crucial and characteristic respects, therefore, the Monetarist approach to events differed from its Keynesian counterpart; first, Monetarists were relatively optimistic about the possibility of deflecting money wages downward from their previous path, and, second, they ascribed much less importance to the effect of incomes on expenditure. To illustrate these differences, we will examine the position taken by Laidler in his evidence to the Expenditure Committee as described in the final report (which was written, it so happened, with the assistance of another Monetarist, Alan Walters, as specialist adviser to the Committee).

On an implicit assumption of flexibility in factor costs, the Monetarist approach[8] proceeded by comparing the "supply" of goods to the home market with the "monetary demand" for them, the latter being related principally to the domestic money stock. For any given level of domestic output, the supply for the home market is defined by what is left over after export demand has been met. As to the level of export demand, the Committee put it to the witnesses that "Realistically they must assume . . . that the full additional cost of

[7] Some of the nondiscretionary increase in government expenditure was the result of public sector wages and salaries rising faster than predicted (over a period in which many such increases were indexed to the retail price index). Some of the adverse cost-inflationary effects of this rise in public sector incomes are considered briefly below.

[8] The approach adopted by *U.K.* Monetarists at that time had little in common with what is now called "the monetary theory of the balance of payments." For more detail of the Monetarist position, reference can be made to the pamphlet by Harry Johnson et al. and especially the Appendix to that piece, which was written by Walters (who had published an earlier version in December 1973).

the increase in the price of oil is borrowed from the suppliers of that oil. Assuming that the same quantity of oil is imported, it follows that the supply of goods on domestic markets is not affected at all by the price rise. The question is then what would happen to the demand for these goods" (para. 48).[9]

Laidler's answer to the question posed was summarized in the next paragraph of the report and we quote this at length:

> Professor Laidler argued that provided there was no increase in the rate of growth of money supply there was no reason to suppose that demand would be further augmented so that there would be no additional inflationary effect. He agreed with all witnesses that there would be an increase in the price of oil and of those goods which use oil intensively relative to the price of other goods. But then in Professor Laidler's terms the rate of inflation of other prices during this period will be somewhat less than the rate of inflation which they would have experienced in the absence of the oil price rise. [para. 49]

So, with the same volume of goods supplied to the domestic market and with no change in the growth of the money supply in response to the change in import prices, we are to expect basically no change in the path of the domestic price index. Some goods will inflate faster, but others less. How about factor incomes? They must, of course, have grown more slowly, because if they showed no change, then the price level would have risen.

The striking difference between the two schools can be graphically illustrated as in Figure 3. In the right-hand panel the principal features of the Keynesian income expenditure treatment are repeated from the previous section. In the left-hand panel, with the price level on the horizontal axis, we portray the essence of the Monetarist account.

We take as a starting point a position of high employment and current account balance with total final sales at $F_0$ and real income at $Y_0^R$. The supply of goods to the domestic market, i.e., for domestic absorption, is total final sales less exports. Together with the initial price level, this "supply" gives the value of domestic expenditure in the initial equilibrium; see point $A$. Through this point is drawn a schedule showing those combinations of domestic absorption and prices which have the same nominal value. Along this schedule, to use Monetarist terminology, "monetary demand" is constant, and as Laidler's account indicates, the level of monetary demand is principally governed by the money stock.

What effect does a change in the terms of trade have *ceteris paribus* on the equilibrium at $A$ where the "demand" and the "supply" schedules intersect? Under the conditions put forward by the Committee, that exports are unchanged and so is the money stock, the answer is, of course, none at all! Hence the strict Monetarist conclusion that with no change in the growth of money, all will be well—no imported inflation and no fall in output or employment.

The Keynesian view was that by appropriate fiscal action, jobs could be preserved, but at the cost of "importing inflation." Hence, with no change in volumes, the value of domestic expenditure would rise as equilibrium shifted (from $A$ to $B$ in the figure) and this would require an appropriately permissive monetary policy. In the absence of fiscal stimulus, however, Keynesians predicted a contraction of aggregate demand which would lower real incomes by more than the fall due to the shift in the terms of trade. In Figure 3 this recession equilibrium is shown at a level of real income labeled $K$, which is associated with a lower level of total final sales ($F_1$) and so a lower supply

[9] The need to borrow overseas is because the exchange rate is assumed to be fixed, and the volume of exports to be constant, and it is for this reason that the supply of goods to the home market is said to be unaffected by the rise in import prices.

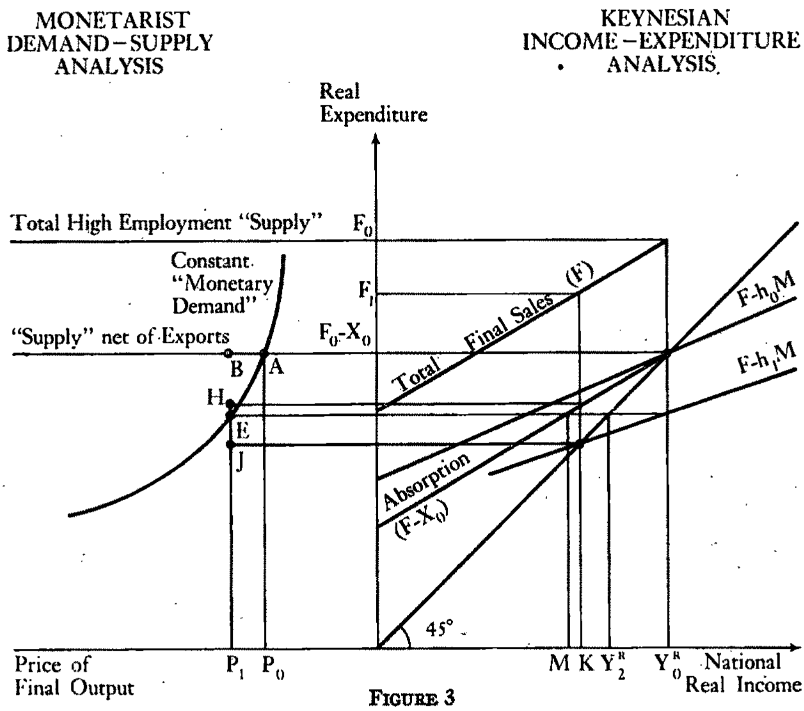



FIGURE 3

of goods to the domestic market. Assuming that prices stay at $P_1$ even in the recession, the value of domestic *expenditure* has risen above its initial level (compare point $H$ with the constant monetary demand schedule), although the value of nominal *income* will have fallen (see point $J$). Since, given a low enough "multiplier," a Keynesian contraction can thus imply a rise in the value of expenditure, a policy of keeping monetary demand constant may require an *even greater* recession. Thus, in the figure, keeping monetary demand constant in the face of prices which do not fall back from $P_1$ would require a contraction of supply to point $E$, which is associated with a fall of real incomes to some point between $K$ and $M$.

So, while Keynesians expected a rise in prices" as an explanation of *U.K.* inflation, and argued for "a gradual return mended expansionary fiscal policy and permissive monetary policy, the Monetarist logic predicted no inflation or recession as a consequence of the change in the terms of trade so long as fiscal and monetary policy were unchanged. Such is the contrast of views and policy recommendations stemming, in my opinion, from the different assumptions as to factor price flexibility and the determinants of aggregate demand mentioned earlier.

It should be mentioned that Laidler did, in his evidence and elsewhere, somewhat qualify the stark Monetarist position portrayed above. Thus he acknowledged that the fall in real income would have some effect on aggregate demand (see House of Commons, para. 187); and he did argue, on the subject of the oil price rise, that "the Government might be wise to permit a temporary inflation in order to help us over these frictions. It *might* be wise and it would have to *permit* it" (see Laidler). It hardly seems likely that Laidler and other Monetarists would wish such qualifications to be taken very far, however, as this would make their position approximate to that of the Keynesians.

However, in early 1974, Johnson, Laidler, Walters, and others signed an open letter to the Prime Minister (see Johnson et al.) which dismissed "the myth of world prices" as an explanation of *U.K.* inflation, and argued for "a gradual return to the aim of a balanced Budget," and recommended cuts in government expenditure and rises in taxation in order to secure a gradual reduction in the rate of increase in the money supply.

In his recent critique of Monetarism, Hicks (1975b) concludes, as I have done, that the Monetarist proposal to prevent imported inflation by controlling aggregate demand would require increased unemployment since domestic factor incomes have to be reduced.[10] "So," he concludes, "a Monetarist policy, applied to the emergency we have been discussing, is strictly deflationary. It involves a contraction of demand, below the level it could have been allowed to reach if the external situation had been more favorable." He goes on to note, "If several countries affected by the shift in the terms of trade all simultaneously adopted the monetarist policy, there is certainly a question whether a contraction, simultaneously adopted in this manner, could be kept under control. It is, therefore, very fortunate that this has not happened. In none of the countries affected has there been an attempt to go so far. But the price of not going so far has been an all-around increase in the rate of inflation" (pp. 8–9).

[10] Members of the Chicago Money and Banking Workshop pointed out that if the demand for nominal balances depended upon nominal *income* ($Y$) (and not nominal domestic *expenditure*, $P(F-X_0)$, as assumed in the above account of the position taken by Laidler and Walters) then a *constant* money stock would in fact "validate" the rise in final output prices due to imported prices. This is the interpretation of the Monetarist position which Hicks followed in his critique. Whether the quantity of money is more stably related to factor incomes or to domestic expenditure on final output appears to be an unsettled issue among Monetarists.

### IV. Real Wage Resistance

The majority of witnesses before the Public Expenditure Committee agreed that the rise in the cost of imported raw materials and oil would increase the price of final output, which would be consistent with nominal wages *not falling* in response to the change in terms of trade. But some witnesses raised the possibility of further rounds of inflation as "people try to offset the increase in their cost of living by asking for more wages" (see comments by Sir Kenneth Berrill in House of Commons report, para. 498). This was put most clearly by G. D. N. Worswick when he argued: "If all of us just took the rise in the price of oil on the chin that would be one thing, but most of us do not; we say, 'Our income is unchanged and prices have risen. We wish to restore our real income.' " (See House of Commons report, para. 136.)

The reason wage earners may resist a fall in the real wages in this manner, according to Hicks (1975b), is that:

> The wage earner's test for fair wages is not simply a matter of comparison with other people's earnings; it is also a matter of comparison with his own experience in the past. It is this which makes him resist a reduction in the purchasing power of his wage, and even a reduction in the growth of that purchasing power to which he has become accustomed." [p. 5]

Hicks has attributed much of the recent *U.K.* inflation to this phenomenon of "real wage resistance."

Now J. D. Sargan (1964) has fitted to *U.K.* data an equation which incorporates this notion that money wage behavior is largely determined by the attempt to restore real wages to a target level which grows steadily through time. Using a modified version of his equations, it is shown in the Appendix how a change in the terms of trade can as a consequence shift the output-inflation tradeoff in the domestic economy.

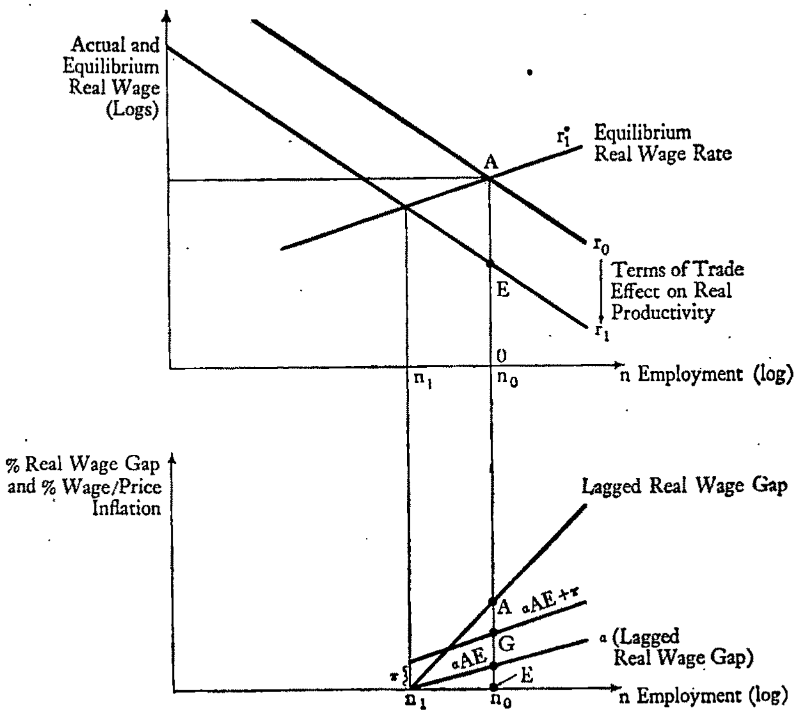



FIGURE 4

The ideas involved are illustrated in Figure 4. The schedule $r_0$ in the top panel shows the marginal product of labor at a point of time as a decreasing function of the number of people employed, $n$, both in *logs*. This marginal product must, in an open economy, be measured net of import requirements. (Thus in the levels it is the addition to the volume of final output *net* of the volume of imports required multiplied by the terms of trade or, to use earlier notation, it is $(dF/dN)(1-mP_I/P)$ where $n=\log N$.) Hence, though a deterioration in the terms of trade will not affect the physical productivity of labor, it will adversely affect the value-added by labor, as shown by the shift from $r_0$ to $r_1$ in the figure. The other schedule $r_1^*$ shows the desired or equilibrium real wage and can be, in the words of Sargan (1964), "interpreted as representing the joint nature of wage-bargaining procedure" (p. 38). The equilibrium real wage was estimated to fall with the number of people employed (relative to a given work force) as the demand for labor and the bargaining strength of the unions declined. It was also reckoned to rise through time (at roughly 2 percent per annum), though there was some evidence that this trend was reduced by the wage freeze at the end of the 1940's. (For simplicity such trends are ignored in the figure.)

The lower panel shows the percentage rate of change of money wages as a fraction $\alpha$ of the percentage "real wage gap" between desired and actual real wages in the period before, *plus* what inflation was then expected. Although inflation expectations were not found to be significant by Sargan (1964, 1971), they are included in this account with a coefficient of one in the light of more recent work on wage determination by J. Johnston and M.

Timbrell.

Of course, if there are no further changes in the exchange rate or foreign prices, then domestic wage/price inflation will in time offset the initial adverse shift in the terms of trade. This, in itself, would have required a considerable rise in British wages and prices after the 1973–74 rise in import prices; but in what follows it seems more appropriate (in the light of what later took place) to assume that the exchange rate depreciates to prevent such domestic inflation reversing the initial change in the terms of trade. Given that there are no subsequent changes in the terms of trade (and ignoring productivity trends), the growth of wages and prices should be the same, and the common rate of wage/price inflation is shown in the bottom panel.

Consider the implications of the initial adverse shift in the terms of trade for inflation in such a model, given that the authorities maintain aggregate demand and "validate" inflation by providing the required money supply. From an initial equilibrium with the real wage at $OA$, and zero inflation (although it could be any other steady-state rate), the system will be moved into disequilibrium with a real wage gap of $AE$, as the level of prices rises relative to wages. As long as there is no change in inflationary expectations, moreover, the rate of wage and price inflation will subsequently be $\alpha\ AE$. If, for example, $\alpha$ is 1/3 and the period is a quarter (see the article by Sargan, 1964), then a 5 percent real wage gap would mean inflation of just under 7 percent per annum. But as is familiar from the debate about the Phillips curve, expectations cannot long remain unchanged in these circumstances, so inflation will tend to accelerate; and, with a unit coefficient on expected inflation with employment at $n_0$ (at point $G$ in Figure 4, for example, actual inflation will exceed what was anticipated).

As can be seen from the figure, the only level of employment where a steady-state rate of inflation is now feasible is at $n_1$. The increase in the level of unemployment required to permit steady-state rates of inflation will depend upon the slope of $r_1$ as well as that of $r_1^*$, but for the latter, Sargan's initial calculations involved doubling the level of unemployment in order to reduce the equilibrium real wage by 3 percent; and subsequent estimates (see Sargan, 1971) have found the equilibrium real wage even less sensitive to the level of unemployment. So the "natural rate of unemployment" implicit in these relationships is quite sensitive to changes in the real wage.[11]

As is evident from the Appendix, this account of inflation as being related to a real wage gap actually bears a strong family resemblance to the "expectations augmented" Phillips curve (differing primarily by the inclusion of the lagged real wage variable). By that token, this explanation is as free of money illusion as the long-run vertical Phillips curve. Thus the nature of equilibrium embodied in the above relationships does not preclude steady-state inflation matched by steady depreciation of the currency and growth in the money stock. Also, as we have noted, under fixed but adjustable exchange rates, a devaluation from a position of initial equilibrium would be offset by domestic inflation, given accommodating monetary policy.

The inclusion of Sargan's lagged real

[11] The natural rate of unemployment here denotes specifically that rate of unemployment associated with the level of employment where the net marginal product of labor equals the bargaining-equilibrium (making steady rates of inflation possible). Note that even if the bargaining-equilibrium real wage at a point of time may be insensitive to the level of unemployment, there is more chance that the *trend* of this equilibrium rate may be lowered by higher unemployment. If this is so the natural rate of unemployment may given time return to the original level, as the net marginal product of labor rises faster than the target real income.

wage variable does make a considerable difference, however, since it explains how there would still be upward pressure on wage costs even if there is no change in the level of employment or in inflationary expectations. It does in this way explain shifts in the natural rate of unemployment (as defined above). These shifts will presumably be reversed in time as the change in the real demand for labor is appreciated by the suppliers of labor, but in the meantime maintenance of high levels of employment will be associated with accelerating inflation.

This description is consistent with Hicks's (1975a) account of the *U.K.* predicament. There he notes that the primary product crisis marked a sharp fall in the real earning power of British labor, and, he continues,

> Now that the veil of money has been so largely stripped off, we have to face it in terms of traditional economics, waking up from the sweet Keynesian dreams that have been with us so long. When the real demand for labor falls (as it has fallen) then, traditional theory says, either there must be a fall in real wages, or there must be unemployment. . . . It is the attempt to bring about such a fall in real wages by a lag in money wages behind money prices—a lag which has by now become quite undependable—that is the principal cause of the acceleration of inflation. [p. 19]

Nor was Hicks optimistic about the capability of fiscal or monetary policy to secure the requisite reduction in the real wage by controlling demand and increasing unemployment. Writing later in the year (1975b), he concluded: "One is therefore led to be more charitable to the 'social contract' and to the '£6 limit'—and to the successors which it is only too likely will have to follow them—than one is tempted to be from one's economist's prejudices" (p. 13).

Unfortunately, however, the operation of Stage 3 of incomes policy over the period November 1973 to July 1974 helped to *offset* any reduction of real wages beneath trend, and so to institutionalize such real wage resistance for a year, as I explain in the next section.

### V. Incomes Policy: Stage 3

Stage 3 was designed under the presumption that the terms of trade would shift in favor of the United Kingdom over the year following October 1973. With wages and salaries subject to a ceiling of 7 percent per annum and productivity growing by 3 percent, it thus appeared reasonable to assume that prices would not rise by as much as 7 percent; hence the policy allowed for compensation to be paid for increases in the price index above 7 percent from the starting date of the policy.

Thus the policy encouraged wage earners to assume that their real incomes would not fall; indeed, since wages were widely expected to "drift" above the 7 percent ceiling up to around 10 percent, the policy was consistent with a rise of 3 percent in real wages (without any triggering of the price threshold payments).

When the terms of trade shifted sharply against the United Kingdom and these higher costs were passed through into the retail price index, the operation of the ceiling on nominal increases had the initial effect of producing the "Keynesian" result that nominal incomes showed no reaction to the price rise and real wages fell. Thus by the first quarter of 1974 real wages had fallen by over 2 percent; see Table 2.

But when three threshold payments were triggered by the Retail Price Index passing the 7 percent mark in April, the situation was transformed completely. For those who had threshold agreements and had settled their annual wage increase, there was no further fall in real wages; for those who had threshold agreements and had yet to settle, the entire

TABLE 2—INCREASES IN AVERAGE REAL WAGES, 1973–75

| Indices 1970=100 | Hourly Wage rates (1) | Retail Prices (2) | Real Wages (1)÷(2) x 100 (3) | Percentage Increase over Quarter: Hourly Wage Rates (4) | Retail Prices (5) | Real Wages (6) |
|---|---|---|---|---|---|---|
| 1973: 1 | 138.4 | 122.2 | 113.3 | — | — | — |
| 2 | 143.6 | 125.3 | 114.6 | 3.8 | 2.5 | 1.3 |
| 3 | 150.2 | 127.7 | 117.6 | 4.6 | 1.9 | 2.7 |
| 4 | 153.7 | 131.8 | 116.6 | 2.3 | 3.2 | −0.9 |
| 1974: 1 | 158.1 | 137.2 | 115.2 | 2.9 | 4.1 | −1.2 |
| 2 | 166.3 | 145.3 | 114.5 | 5.2 | 5.9 | −0.7 |
| 3 | 181.2 | 149.4 | 121.3 | 9.0 | 2.8 | 6.2 |
| 4 | 193.0 | 156.1 | 123.6 | 6.5 | 4.5 | 2.0 |
| 1975: 1 | 205.7 | 165.6 | 124.2 | 13.3 | 6.1 | 7.2 |
| 2 | 221.8 | 180.7 | 122.7 | 7.8 | 9.1 | −1.3 |
| 3 | 235.0 | 188.7 | 124.5 | 6.0 | 4.4 | 1.6 |

*Source:* for cols. (1) and (2): *NIESR Review*, various issuses.

money wage increase of the settlement was fully indexed. Assuming roughly half the settlements were still to come in the second half of the year following September 1973, an average fully indexed money wage increase of about 10 percent per settlement would raise the national average wage by about 5 percent, which would then offset the 2 percent fall in real wage over the preceding six months to give a rise of 3 percent over the year as a whole. As can be seen from Table 2, real wages did in fact rise by just over 3 percent from third quarter 1973 to third quarter 1974.

In terms of the analysis of the last section, the situation under the incomes policy after April was one of very powerful "real wage resistance" which could be expected to produce a very rapid acceleration of inflation as prices were marked up on wages and so triggered further price increases. This did, of course, happen to some extent, as the 5 percent rise in real wages mentioned above was achieved by a 14 percent rise in wages relative to a 9 percent rise in prices in six months; see Table 2 and Figure 1.

The acceleration of inflation required to *reduce* real wages, under indexation which adjusted wages to prices with a delay of about a month or so, would have been enormous, and as we have seen real wages actually *rose on trend* over the year from 1973 third quarter. This rise was made possible partly by increasing subsidization of prices by the government and partly by a squeeze on company profits as prices were held back under the provision of the Price Code. According to the National Institute (*NIESR Review*, 1975, p. 31) the operation of this Code probably reduced the price level by something of the order of 2 percent below what it might otherwise have been by the third quarter of 1974. The increase in the level of subsidies was by no means sufficient to prevent the real wage rise from reducing profits, and the National Institute (*NIESR Review*, 1975, p. 35) noted that the share of profits (after stock appreciation and depreciation) fell from 12.6 percent of national income in 1973 to 10 percent in third quarter of 1974 (and the Financial Times index of share prices fell by roughly one half over the same interval).

It is apparent that the Government made an error in promoting the indexation of wages to the Retail Price Index (which of course includes import prices and indirect taxes as well as "home costs").[12]

There were those who suggested therefore that the government should shoulder the responsibility of actually *paying* the threshold, giving as it were a wage subsidy, the costs of which could have been borrowed from abroad. Such a scheme, like that discussed by John Flemming would have acted to prevent the cost of labor from rising above its productivity. (Thus in terms of Figure 4, a subsidy of $AE$ per capita could have preserved real wages constant and prevented the profit share from falling.) The assumption behind such proposals was that real wages would not rise (above the level given by the initial subsidy) over a period of a year or more, so the rising productivity of labor would reduce the subsidy required. But subsequent experience has shown that any such subsidy could only have been paid as part of a tough incomes policy. Under the voluntary "Social Contract" prevailing after the end of Stage 3, real wages, far from falling below trend, rose very rapidly, increasing by 7.2 percent in the first quarter of 1975 alone!

It is clear that after Stage 3 real wage resistance as described by Hicks was overtaken by a more determined upward pressure for real wages. Partly, no doubt, labor leaders were encouraged by their ability to negotiate rapidly rising real wages earlier that year, but there was also a lack of will on the part of the Labour government to resist wage rises in the public sector (following the experience of the Conservative Administration with the miners). The transmission of these wage pressures to the private sector was aided by talk of "rescuing" or nationalizing companies which might be unable to pay such increases. Another factor reckoned to have put upward pressure on real wages was the increasing burden of taxation required to pay for the price subsidies, transfers, and spending programs of the Labour government. The Chancellor of the Exchequer appealed to labor unions to take account of the value of these various components of government *spending* when they set their wage demands (arguing that these elements, dubbed the "Social Wage," were worth £20 per week —over a third of the national average wage).

In spite of the Chancellor's appeals, it was increasingly evident that money wage settlements were increased in response to higher taxation *despite* the benefits flowing from the expenditure financed therewith. Dudley Jackson, Herbert Turner, and Frank Wilkinson had in 1972 argued that there existed a wage-tax spiral, and econometric evidence that taxes affected wage settlements was provided by Johnston and Timbrell in 1973.

It thus appeared possible, alarmingly enough, that an increase in taxation could have much the same sort of effect on the inflation-employment relationship as the rise in the price of imports examined in the last section. This would mean that the maintenance of the same level of output and employment by aggregate demand management would be associated with increasing inflation. Preventing wage inflation from raising prices by price controls, moreover, would simply lead to a collapse of profits and so investment. In the article already cited, Johnston and Timbrell noted that the "perverse" positive relationship between wage inflation and unemployment they found there "would be consistent with a theory that suggested variations in unemployment as the *consequence* of changes in wage rates, so that increases in money wages, beyond what current shifts in the production function

[12] The case for choosing the $GDP$ deflator as a more appropriate index for wage contracts in these circumstances is made by Jackman and Kurt Klappholz. In this they depart from what Milton Friedman had to say when he spoke about $U.K.$ inflation on B.B.C. radio soon after the oil price rise and appeared to endorse index linking to a general price index, as under the threshold arrangement.

and current market conditions . . . would bear, result in lower employment and higher prices" (p. 165). Such a consequence has indeed followed the rapid rise in real wages in 1974 and 1975, when it was found that merely maintaining aggregate demand was not sufficient to ensure the employment of labor whose cost to employers exceeded its productivity as increasing labor costs outstripped price inflation.

## VI. Conclusion

I conclude first that it was virtually inevitable that the steep rise in the price of imports from 1973 to 1974 should have resulted in a sharp rise in domestic prices. In this, I am in agreement with the majority of witnesses giving evidence before the Public Expenditure Committee. Furthermore, so long as wage and salary earners were unwilling to reduce their real income beneath the path they would have followed without the shift in the terms of trade, there existed pressures to perpetuate and amplify this "first round" of price increases. Such latent real wage resistance was undoubtedly strengthened by the prevailing incomes policy which failed to allow for the need to reduce real incomes and encouraged the use of threshold agreements which linked money wages to the Retail Price Index.

How much of the increase in money wages was a consequence of the provisions of the incomes policy (which involved eleven "triggers" of threshold payments in seven months) is, as ever, a matter of assessing what would have happened in their absence; doubtless there would have been pressure for interim pay awards, but I conclude that the incomes policy added significantly to the backlash of prices on wages. If incomes policy can in this way affect wage behavior for the worse, it can also in my view affect it for the better (see, for example, the effect of incomes policy in July 1975 after the real wage explosion under the voluntary social contract).

In the circumstances described, "Keynesian" policy recommendations which focused simply on maintaining the level of aggregate *real* demand failed to foresee the collapse of private investment demand and employment as rising real wages made it increasingly unprofitable to "supply" the goods demanded. The Monestarists' prescriptions for controlling *nominal* demand were, in my opinion, based on an excessively optimistic view as to the flexibility of real and nominal wages; consequently a deeper recession would have been required for Monetarist policy to work in the circumstances than they would have expected.

While the recent inflation in the United Kingdom is hardly something of which one can be proud, I end by noting some benefits which have flowed from the experience. First of all, it has made union leaders more aware of the vulnerability of investment and employment in the private sector to real wage pressures, so that it is now more widely appreciated that large wage increases that do not lead to price increases will probably lead to unemployment. Second, the reaction of money wages to increasing taxation has impressed on politicians the need to weigh against the political benefits of increased government expenditure the political costs of such inflationary reactions to the taxes required to finance them. Finally, it has made the public at large afraid of inflation and more willing to do what is necessary to prevent it.

## Appendix

### 1. *Keynesian Demand Management*

Using *GDP* at current factor cost as a measure of "nominal income," $Y^N$, implies

$$Y^N = PF - P_I M$$

where $F$ is total final sales, $M$ the volume of imports, $P_I$ the price of imports, and $P$ the price of total final sales (*net* of indirect taxes). For simplicity, it is assumed that the

price of exports moves in line with the price index for total domestic expenditure. "Real income" $Y^R$ may be obtained by deflating $Y^N$ by the price of final output, so

$$Y^R \equiv \frac{Y^N}{P} = F - \frac{P_I}{P} M = F - hM$$

where $h$ denotes the ratio of the price of imports to that of domestic output and is thus a measure of the terms of trade.

The most common behavioral assumptions can be summarized as

$$\begin{aligned} M &= mF \\ F &= C + I + G + X \\ &= C(Y^R - T) + I + G + X \end{aligned}$$

using standard notation for the components of expenditure, and letting $T$ denote the real level of direct and indirect taxes. To maintain total final sales constant by demand management in the face of changes in the terms of trade, the appropriate fiscal policies are implied by

$$dF = C'dY^R - C'dT + dG + dX = 0$$

assuming no change in investment (and no relative price elasticity of imports, which is not too inappropriate for raw material and energy imports). The change in real income if total final sales are constant is simply $-mF\ dh$; so if the volume of exports does not respond with any alacrity, the requisite fiscal policy actions on this Keynesian analysis are:

$$\begin{aligned} dG\,\big|_{dT=0} &= -C'dY^R = C'mFdh \\ dT\,\big|_{dG=0} &= dY^R = -mFdh \end{aligned}$$

Splitting taxes between indirect and direct taxes as follows:

$$T = \frac{t_i}{1 + t_i}(F - X) + T_v$$

where $t_i$ is the rate of indirect taxation on total domestic sales, the appropriate change in the indirect tax rate, *ceteris paribus*, is

$$\frac{dt_i}{(1 + t_i)^2} = \frac{mF}{F - X} dh$$

The dependence of taxes on income is ignored in Figure 2 of the text.

2. *Inflation, Real Wages, and Employment*

The following is a modified version of Sargan's equations:

*Target or Bargaining Equilibrium Real Wage*

(A1) $\quad r^* = \beta_0 + \beta_1 n + \beta_2(\mathrm{t} + 1)$

*Expectations-Augmented Wage Adjustment Equation*

(A2) $\quad \Delta w = \alpha(Lr^* - Lw + Lp) + L\pi$

*The Price Equation*

(A3) $\quad p = \gamma_0 + \gamma_1 p_I + (w - R)(1 - \gamma_1) + \gamma_2 T_i$

*The Determination of Output per Capita*

(A4) $\quad R = \gamma_3 + \gamma_4(k - n) + \gamma \mathrm{t}$

where $w = \log$ of index of money wage rate
$p = \log$ of index of retail prices
$p_I = \log$ of index of import prices
$R = \log$ of index of output per capita
$T_i = \log$ of ratio of consumption expenditure at market prices to consumption expenditure at factor cost
$r^* = \log$ of equilibrium or target real wage
$n = \log$ of number of people employed
$k = \log$ of capital stock
$\pi =$ anticipated inflation
$\mathrm{t} =$ time, measured in quarters from some base date
$L =$ the lag operator, $p(\mathrm{t}-1) \equiv Lp(\mathrm{t})$
$\Delta \equiv 1 - L$

Assuming that for $\mathrm{t} > 0$ there is no change in indirect taxes, employment, capital stock, or the relative price of imports and domestic output (as when the exchange rate floats to offset domestic inflation), then we have $\Delta p = \Delta p_I$, $\Delta T_i = 0$, $\Delta R = \gamma$, $\Delta r^* = \beta_2$ for $\mathrm{t} > 0$ and so, from (A3) and (A4),

(A5) $\quad \Delta p = \Delta w - \gamma$

which, if we denote the actual real wage by $r$, can be rewritten as

(A6) $\quad \Delta r = \gamma$

Substitution into (A2) yields

(A7) $\Delta p = \alpha(Lr^* - Lr) - \gamma + L\pi$

so the rate of inflation increases with the lagged gap between the equilibrium real wage $r^*$ and that which is actually paid $r$, and also with lagged inflation expectations $\pi$. (The net negative effect of productivity growth is shown in (A9) below.) Corresponding to (A6) we can describe the *level* of real wages $r(t)$ as

(A8) $r(t) = r(0) + \gamma$

where from (A3) and (A4) we find that the real wage at time zero is $r(0) = \gamma_1(w(0) - p_I(0)) + (1-\gamma_1)(\gamma_3 + \gamma_4(k(0) - n(0))) - \gamma_0 - \gamma_2 T_i(0)$ So $r(0)$ is a decreasing function of $n(0)$, as in Figure 4. By substituting into (A7) we obtain

(A9) $$\Delta p = \alpha(\beta_0 + \beta_1 n(0) - r(0)) - (1-\alpha)\gamma + L\pi$$

on the assumption that the trend in the equilibrium real wage $\beta_2$, matches the trend in output per man $\gamma$. Since $n$ will vary inversely with unemployment, this equation is not unlike an "expectations-augmented Phillips curve," modified to take account of the level of real wages. Assuming that no inflation is expected, then the rate of inflation is simply given by

(A10) $$\Delta p = \alpha(\beta_0 + \beta_1 n(0) - r(0)) - (1-\alpha)\gamma$$

which is the rate discussed in the text; but if $\pi = \phi(L)\Delta p$ where $\phi(1) = 1$, then (A9) implies that inflation will accelerate upwards or downwards depending on the sign of the right-hand side of (A10).

Thus, one could by setting (A10) equal to zero and "solving" for $n(0)$ define a "natural rate of employment" (and correspondingly a "natural rate of unemployment") which would prevent any *changes* in the rate of inflation and so permit expectations defined by $\phi(L)\Delta p$ to be fulfilled. This natural rate of employment will fall if $p_I(0)$ rises, see Figure 4.

Indirect taxes have the effect of lowering the real wage as defined above. To allow for the additional possibility discussed in the text that the speed of adjustment of money wages depends upon the gap between a desired real wage and the real wage *net of income tax*, and assuming for simplicity a proportional rate of income tax, one could modify equation (A7) to read:

(A11) $\Delta p = \alpha(Lr^* - Lr - LT_x)$

where $T_x = log$ (1 − proportional rate of income tax). Assuming the tax rate is unchanged t = 0, this would change (9) to

(A12) $$\Delta p = \alpha(\beta_0 + \beta_1 n(0) - r(0) - T_x(0)) - (1-\alpha)\gamma + L\pi$$

where $r(0)$ is defined as before. It is obvious from this how a rise in rate of income taxation has the same impact on the inflationary process as a reduction in real wages by a change in the terms of trade.

## REFERENCES

J. **Flemming,** "Adjust the Real Elements in a Changing Economy," *Catch '76..?,* Occas. Paper 47, Inst. Econ. Affairs, London 1976, 9–16.

J. **Hibbert,** "Measuring Changes in the Nation's Real Income," *Economic Trends,* Jan. 1975, *255,* XXVIII-XXXV.

J. R. **Hicks,** (1975a) "The Permissive Economy," *Crisis '75..?,* Occas. Paper 43, Inst. Econ. Affairs, London 1975, 17–21.

———, (1975b) "What is Wrong with Monetarism," *Lloyds Bank Rev.,* Oct. 1975, 1–13.

R. **Jackman and K. Klappholz,** "The Case for Indexing Wages and Salaries," in T. Leisner and M. King, eds., *Indexing for Inflation,* London 1975, 19–40.

D. **Jackson, H. A. Turner, and F. Wilkinson,** "Do Trade Unions Cause Inflation?," Occas. Paper 36, Univ. of Cambridge 1972.

H. G. **Johnson et al.,** *Dear Prime Minister,* London 1974.

J. **Johnston and M. C. Timbrell,** "Empirical Tests of a Bargaining Theory of Wage Rate Determination," *Manchester Sch. Econ. Soc. Stud.,* June 1973, *41,* 141–67.

D. **Laidler,** "Discussion" in Robbins et al., eds., *Inflation: Causes, Consequences, Cures; Readings,* no. *14,* London 1975.

J. D. **Sargan,** "Wages and Prices in the United Kingdom," in P. E. Hart et al., eds., *Econometric Analysis for National Economic Planning,* London 1964.

———, "A Study of Wages and Prices in the U.K.: 1949–1968," in H. G. Johnson and A. R. Nobay, eds., *The Current Inflation*, London 1971.

**A. A. Walters**, "Importing and Exporting Inflation," *Int. Currency Rev.*, Nov./Dec. 1973, *63*, 7–10.

**House of Commons**, Ninth Report from the Expenditure Committee, HC-328, Session 1974–1975: "Public Expenditure, Inflation and the Balance of Payments," London 1974.

**National Institute of Economics and Social Research**, *National Institute Economic Review* (*NIESR Review*), "The Home Economy," May 1974, *68*, 9–44.

———, "The Economy in 1974," Feb. 1975, *71*, 8–39.

# The British Inflation: Indigenous or Imported?

*By* JOHN WILLIAMSON AND GEOFFREY E. WOOD*

Since 1970 it has been asserted with increasing force that the recent worldwide acceleration of inflation has a common cause (see, for example, Hans Genberg and Alexander Swoboda; David Laidler and Michael Parkin, Section 5; William Nordhaus). This cause is usually found in the expansionary monetary policies pursued in the United States during the period of the Vietnam War. Our principal purpose in this paper is to examine whether this theory provides a convincing explanation for the acceleration of inflation in the United Kingdom in the period since 1967, which would imply that Britain's inflationary problems have in large measure been imposed upon her.

One possible basis for contending that British inflation was caused by global inflation lies in the view that arbitrage in traded goods is sufficiently perfect to prevent the price of British-produced tradeables diverging significantly from the price level set on the world market. We examine, and reject, this view in Section I. This means that British inflation cannot be explained as merely a local reflection of worldwide experience, but needs to be explained in terms of both the operation of the British economy and its reactions to developments in the rest of the world. In Section II we therefore attempt such an explanation on the basis of an examination of recent British economic history. In Section III we state our conclusions about British experience and sketch some implications for the interpretation of global inflation.

* University of Warwick and City University, London.

## I. The Perfect Arbitrage Hypothesis

The hypothesis that the "law of one price" prevents the price of tradeable goods produced in different countries diverging even in the short run has been popularized in many of the simpler monetarist models of the balance of payments, and forms the basis of the "Scandinavian model" of inflation (see, for example Gösta Edgren, Karl-Olaf Faxén, and Clas-Erik Odhner). This hypothesis is to be sharply distinguished from the hypothesis that in the long run the inflation rates in different countries which maintain fixed exchange rates between themselves must move in parallel; the latter can hold, even if short-run substitution elasticities are relatively low, through the income effects and monetary repercussions of trade imbalances which are ultimately induced by the emergence of price discrepancies. The hypothesis of no short-run price independence rests on the presence of very high substitutability between the goods produced in different countries. This is plausible enough where primary products are concerned (and some empirical support has been provided by Genberg, pp. 21–23), but questionable in the case of manufactured goods, which is the case of overwhelming importance for a country such as the United Kingdom.

Table 1 shows three sets of comparisons between British and foreign prices of manufactures. Column (1) compares the unit value index for British exports with that for all countries, while column (2) provides a similar comparison with the *U.K.*'s principal individual trading competitor. Column (3) shows a similar comparison be-

TABLE 1—RELATIVE PRICES OF BRITISH AND FOREIGN MANUFACTURED GOODS, 1966–75
(1970=100)

| | Ratio of British export prices to foreign export prices (1)[a] | Ratio of British export prices to German export prices (2)[b] | Ratio of British wholesale prices to British import prices (3)[c] | Effective exchange rate (4)[d] | *DM/£* exchange rate (5)[e] |
|---|---|---|---|---|---|
| 1966 | 107 | 105 | 111 | 115 | 126 |
| 1967 | 104 | 106 | 111 | 114 | 126 |
| 1968 | 99 | 102 | 99 | 100 | 109 |
| 1969 | 101 | 101 | 98 | 100 | 107 |
| 1970 | 100 | 100 | 100 | 100 | 100 |
| 1971 | 101 | 98 | 108 | 100 | 97 |
| 1972 | 103 | 99 | 110 | 95 | 91 |
| 1973 | 95 | 86 | 98 | 85 | 75 |
| 1974 | 94 | 89 | 91 | 82 | 69 |
| 1975 | 96 | 91 | 99 | 74 | 62 |

[a] Unit value index for exports of British manufactures (SITC 5–8), divided by the unit value index for manufactured exports of all countries. *Source: NIESR Review*, Table 19.
[b] As for column (1) except that the unit value index for German manufactured exports replaces that for all countries.
[c] The British wholesale price index for home sales divided by the unit value index for British imports of manufactures. *Source: Monthly Digest of Statistics*, Tables 141, 173.
[d] Bank of England *Quarterly Bulletin*, Table 29, and calculations extrapolating effective rate calculations back prior to 1972, on the basis described in *Economic Trends*, June 1974.
[e] *DM* per pound from trade conversion factors in *International Financial Statistics*.

tween British wholesale prices and the unit value index for imports of all manufactures. All three columns show a degree of variation, with ranges from 13 to 23 percent, that hardly seems consistent with perfect arbitrage. Comparison with columns (4) and (5) suggests that these variations are in part explicable as a response to exchange-rate changes, which, according to the perfect arbitrage hypothesis, should have no impact on relative prices.

It is of course true that unit value indices leave something to be desired as price indices because changes in the quantity composition of trade can cause a change in the index even if no individual price is altered. There is little reason to dismiss the evidence exhibited in Table 1 on this ground, however, since the only systematic bias introduced is that price divergencies are *under*stated inasmuch as goods that become too expensive to sell tend to drop out of the index. Nevertheless, both in order to provide a check on the aggregate comparisons shown in Table 1 and because the subject is of inherent interest, an examination was made of micro price changes in what may plausibly be considered the representative manufactured tradeable good, the automobile. This was done by calculating annual price changes for individual automobile models produced in the United Kingdom and models imported into the United Kingdom from each of the five countries (France, Germany, Italy, Japan, United States) which exported four or more types of automobile to the United Kingdom during the period 1967–75. The outstanding characteristic of these figures was the range of the price increases recorded within each year; the standard deviation of price increases averaged 6.8 percent with a minimum of 4.6 percent, while the range averaged 38.5 percent with a minimum of 25.1 percent. (For British-produced cars the minimum standard deviation was 3.8 percent and the minimum range was 17.4 percent.) These variations in relative prices are inconsistent with the view that there is perfect

TABLE 2—FREQUENCY OF PRICE INCREASE OF FOREIGN-PRODUCED AUTOMOBILES IN *U.K.* MARKET, 1967–75

| | Sterling depreciation greater than 5 percent | Sterling appreciation greater than 5 percent | No "significant" exchange-rate change | Total |
|---|---|---|---|---|
| Range of price increase in terms of price increases of British-produced models | | | | |
| Bottom quartile or below | 11.5 | 8 | 67 | 86.5 |
| Second quartile | 24 | 2 | 30 | 56 |
| Third quartile | 19.5 | 1 | 7 | 27.5 |
| Top quartile or above | 31 | 0 | 13 | 44 |

*Note:* Price data were the manufacturers' recommended retail prices in the United Kingdom, including tax, given in the first issue in May each year of *Motor*. One figure (the first listed) was selected for each maker which supplied identically described models in consecutive years. This procedure gave between 21 and 27 observations for British-produced cars, and between 25 and 28 observations for foreign-produced cars in each year.

arbitrage, and suggest rather that automobiles have to be analyzed as differentiated products whose market conditions permit differential price movements.

These data also permitted a further test of the prediction of the perfect arbitrage hypothesis that prices are unaffected by exchange-rate changes. The observations on the price increases of British-produced models were used to calculate medians and quartiles for the price increases in each year. These figures were then used to classify price increases of foreign models into the four categories shown in the rows of Table 2. The observations of foreign model price increases were also classified into years in which sterling depreciated by more than 5 percent against the currency of the producing country (col. 1), those in which sterling appreciated by more than 5 percent (col. 2), and the remainder, in which there was no significant exchange-rate change (col. 3). The perfect arbitrage hypothesis would suggest that the distributions recorded in the three columns should not differ significantly, while the more traditional view that exchange-rate changes produce at least short-run effects on price competitiveness would lead one to expect a preponderance of observations in the bottom row of column (1) and the top row of column (2), relative to column (3). A *chi*-square test reveals that the tendency in this direction shown in Table 2 is not consistent with the view that the three columns are drawn from a common population.[1] In view of the absence of any attempt to allow for other factors, such as differential inflation, that could in general be expected to offset at least in part the impact of exchange-rate changes, this constitutes decisive refutation of the view that changes in exchange rates do not produce a short-run impact on prices. It follows that a more subtle explanation of the international interdependence of inflation rates is required than the hypothesis that they are rigidly linked together by goods arbitrage.

## II. External Factors in the British Inflation

The analytical framework which we regard as appropriate for analyzing inflation is in most essential respects the same as that endorsed by Laidler and Parkin in Sections 3 and 4 of their admirable survey of inflation theory. Inflation has to be analyzed within a comprehensive macroeconomic model encompassing a theory of

[1] The value of *chi*-square was 55.2 with six degrees of freedom, as against a critical value of 12.6 at the 5 percent confidence level.

TABLE 3—BRITISH INFLATION AND MEASURES OF EXTERNAL INFLUENCES

(All figures are percentages)

| | Wage increase | Increase of final output prices | Inflation expectations | Unemployment rate | Real trade surplus (percent of GNP) | Increase in money supply ($M_3$) | Impact of reserve increase on money supply | Effect of import price increase on real money supply | Increase in price of nonmanufactured imports | Increase in price of manufactured imports | Increase in price of all imports | Effect of terms-of-trade change on potential consumption | Sterling depreciation |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) | (13) |
| 1960–65 average | 5.9 | 3.1 | 3.1 | 1.62 | −0.6 | 6.7 | −4.5 | − 0.2 | 0 | 1 | 1 | \|0.4\| | 0 |
| 1966 | 7.7 | 4.1 | 4.2 | 1.40 | +0.4 | 3.6 | −1.9 | − 0.3 | −1 | 4 | 2 | +0.4 | 0 |
| 1967 | 4.1 | 2.7 | 5.9 | 2.23 | −1.5 | 9.7 | −4.0 | 0 | 0 | 1 | 0 | +0.2 | 1 |
| 1968 | 7.8 | 5.6 | 6.9 | 2.37 | −1.2 | 7.6 | −9.6 | − 2.0 | 8 | 17 | 11 | −0.7 | 14 |
| 1969 | 7.7 | 5.0 | 6.1 | 2.34 | +0.2 | 3.2 | 4.6 | − 0.9 | 5 | 5 | 5 | −0.2 | 0 |
| 1970 | 13.9 | 7.3 | 7.5 | 2.52 | 0 | 9.4 | 8.4 | − 0.7 | 3 | 4 | 4 | +0.8 | 0 |
| 1971 | 13.2 | 8.0 | 7.4 | 3.29 | +0.5 | 11.9 | 11.9 | − 0.9 | 8 | 1 | 5 | +0.2 | 0 |
| 1972 | 13.3 | 6.8 | 5.7 | 3.63 | −1.5 | 23.9 | 0.2 | − 0.7 | 5 | 3 | 4 | 0 | 5 |
| 1973 | 11.0 | 10.4 | 6.9 | 2.55 | −1.4 | 26.2 | −3.8 | − 6.1 | 33 | 21 | 27 | −3.1 | 10 |
| 1974 | 20.2 | 19.0 | n.a. | 2.53 | 0 | 12.3 | −4.7 | −15.6 | 71 | 33 | 53 | −6.1 | 4 |
| 1975 | 27.0 | 23.7 | n.a. | 3.87 | +1.1 | 7.5 | −2.7 | − 3.2 | 11 | 13 | 12 | +2.8 | 10 |

*Note:* All figures are percentages; n.a. = not available.
Col. (1): Year-on-year increase in hourly earnings in manufacturing. *Source: NIESR Review*, Table 8.
Col. (2): Year-on-year increase in total final prices. *Source: N.I.E.S.R. Review*, Table 7.
Col. (3): Estimate of inflation expectations, based on survey data of John Carlson and Parkin, pp. 136–37. Average of monthly figures.
Col. (4): *Source: NIESR Review*, Table 6.
Col. (5): The trade surplus that would have been realized on the basis of 1970 prices and actual trade volumes, as a percent of *GNP*. *Source: Economic Trends*, annual supplement, Dec. 1975, pp. 97, 102.
Col. (6): $M_3$, sum of seasonally adjusted end-quarter figures. Series starts 1963 (first row is therefore 1963–65 average). *Source: Statistical Abstract*, Table 12.2.
Col. (7): The percentage change in $M_3$ resulting from reserve flows (including reserve borrowing). First row is 1963–65 average. *Source: Statistical Abstract*, Tables 12 and 21.
Col. (8): Negative of col. (10), multiplied by the ratio of imports to *GNP*. This provides an estimate of the impact effect of import-price increases on real money balances, excluding any arbitrage-induced rises in domestic prices.
Col. (9): Change in the unit value of imports of SITC 0–4. *Source: Monthly Digest of Statistics* (*MDS*) Table 141.
Col. (10): Change in the unit value of imports of SITC 5–8. *Source: MDS*, Table 141.
Col. (11): Change in the unit value of imports of SITC 0–8. *Source: MDS*, Table 141.
Col. (12): Change in the terms of trade, multiplied by the average of visible exports and imports, divided by private consumption. *Source: MDS* Tables 1, 124, and *NIESR Review*, Table 13.
Col. (13): Depreciation of effective sterling exchange rate. *Source:* see col. (4) Table 1.

the determinants of aggregate demand, a theory of the wage and price determination process, and a theory of the determination of the relevant expectations. Such a framework provides a wide variety of channels through which external factors can impinge on the inflationary process. Table 3 presents data designed to outline the inflationary experience in Britain, and to quantify the major channels through which a comprehensive theory would suggest external factors could have exerted an influence on British inflation.

There is no great difficulty in accounting for British inflation up to the middle of 1969. On average the economy was run at a very high pressure of demand (see col. 4): there existed a rather general consensus that unemployment in excess of $1\frac{3}{4}$ percent —or, to the conservative, $2\frac{1}{4}$ percent—signified a need for prompt demand reflation. The Phillips curve seemed a reasonably well-determined relationship which explained why such an ambitious unemployment target would produce marginally more inflation than was compatible with the maintenance of international price competitiveness, and in fact British prices rose about 1 percent per annum faster than the international norm. There was also, however, a widespread belief that the experience of 1948–50—when government-inspired wage restraint was credited with having produced two observations substantially below the Phillips curve, despite a major devaluation—demonstrated that

the Phillips curve was not an immutable relationship, and this led to hopes that incomes policies might enable inflation to be slowed to the international norm without modifying the employment target, which was near universally regarded as a priority commitment. In the event these hopes were not fulfilled.[2] Export prices became increasingly uncompetitive, and hence the employment target became increasingly incompatible with a satisfactory payments outcome. This process could not go on indefinitely, and was eventually terminated by the devaluation of the pound in 1967.

The immediate and almost mechanical effect of devaluation was to increase import prices, which rose in proportion to devaluation in the case of manufactured imports (col. 10) but substantially less in the case of nonmanufactured imports (col. 9). Since the ratio of imports to *GNP* is about 0.2, the average rise in import prices of 11 percent in 1968 provides a simple cost-push explanation for the higher rate of price inflation in 1968–69 as compared to 1966—a year in which the wage increase was similar (cols. 1 and 2). This initial and modest acceleration of inflation can be regarded as the reimportation of the inflation that had been exported during the years 1960–67, when there was a rise in British export prices some 6 percent greater than the international norm, a cumulative current account deficit of some $2.6 billion, and a cumulative reserve loss of $4.7 billion (which was financed entirely by reserve borrowing). During the years 1968–70, price competitiveness was restored to the level of the early 1960's, the current account surplus totalled some $2.2 billion and the net gain in reserves amounted to $2.1 billion (despite the substantial reserve loss in the first postdevaluation year, 1968).

It is worth pausing at this stage to examine whether, as has often been argued (for example, by Edward Bernstein), the maintenance of a pegged exchange rate prior to 1967 performed the function of making the foreign sector a built-in stabilizer, thus alleviating demand pressures at cyclical peaks and so mitigating inflation. There is a general belief in Britain that this has been the case, although it is usually described as the "problem" of how the balance of payments goes into deficit as full employment is approached. Comparison between the trade balance as usually expressed, in nominal terms, and the state of the cycle suggests that there was some tendency in this direction. For example, the correlation coefficient between year-to-year changes in the trade balance as a percent of *GNP* and the unemployment rate for the four quarters starting in the middle of the year,[3] was 0.62 over the period 1960–67. However, it is not the change in trade balance in nominal terms, but rather its change in real terms that is relevant from the standpoint of providing a built-in stabilizer. The correlation coefficient between the real change in the trade balance (as is shown in col. 5) and unemployment on the same basis over the same period was only 0.39. One must therefore conclude that the existence of a pegged exchange rate provided only a modest alleviation of inflationary pressure through induced changes in the trade balance. The explanation of this surprising result presumably lies in the very high degree of synchronization between the business cycle in the United Kingdom and that in the rest of the world.

Even before the second half of 1969 there were some indications that inflationary pressures were greater than could

[2] While quite a number of the studies that have been conducted on the effectiveness of incomes policies have found at least temporary effects on the rate of inflation from some periods of incomes policy, the particular policies identified as successful have an embarrassing tendency to vary between authors. In addition to this lack of robustness, none of the reported estimates are large.

[3] The time displacement was introduced to allow for the fact that unemployment is a lagging indicator of demand pressure.

be accounted for by the near mechanical cost-push effects of devaluation, in that existing estimates of the Phillips curve predicted lower rate of wage increase than those that actually occurred. In retrospect there are three plausible explanations of this development. First, there were major improvements in the social benefits available to the unemployed in 1966. Second, inflationary expectations were substantially raised by devaluation (col. 3: monthly figures show a sharp jump to 11.9 percent in November 1967, which fell back subsequently, but not to predevaluation levels). Third, the lesser pressure from foreign competition after devaluation presumably increased the willingness of firms to concede wage increases.

However, the major acceleration of British inflation occurred in late 1969, when the first "wage explosion" got under way with conspicuous wage settlements in excess of 15 percent. In the following year wage increases were some 6 percent higher than the 7 to 8 percent range that had previously been the maximum (col. 1). (It seems clear enough that wage increases, rather than price increases, were the initiating factor in the intensification of inflation: the contrary interpretation is inconsistent with the severe squeeze on profits of recent years.)

It is this first wage explosion in 1969–70 which has always posed the greatest challenge to accepted theories of inflation, as expounded by Laidler and Parkin, and hence led many of those who do not question the importance of demand and expectational factors in generating inflation to believe that these do not exhaust the list of important independent influences on the wage-determination process. Unemployment was high by previous, if not by contemporary, standards, and rising. This slack in the economy arose from the impact of the deflationary fiscal policy adopted in 1968, and despite the major diversion of resources into the balance of payments (equivalent in real terms to some $1\frac{3}{4}$ percent of *GNP* as compared to 1967); deliberate fiscal deflation was, incidentally, by far the most important method whereby domestic absorption was curtailed to make room for a payments improvement, vastly outweighing the deflationary impact of devaluation on real money balances which receives so much attention in the monetarist analysis of devaluation (col. 8). Price inflation, though still above the predevaluation level, had declined somewhat from the level reached under the immediate impact of devaluation, which would seem to preclude an adaptive-expectations explanation based on observation of changes in the price level; and this conclusion is confirmed by the decline in the Carlson-Parkin measure of consumer price expectations, which is based on survey data (col. 3). Monetary policy, under the tutelage of the International Monetary Fund (*IMF*), was tight—the money supply grew by only 3.2 percent, and domestic credit expansion was actually negative; which would seem to preclude a rational-expectations explanation based on the observation of monetary aggregates. Hence none of the factors emphasized in the standard theory of inflation in a closed economy offers an explanation for this first wage explosion.[4]

[4] It is only proper to acknowledge that the latest wage equations of the Manchester Inflation Programme, as in the recent paper of Malcolm Gray, Parkin, and Michael Sumner, do not suffer from embarrassing positive residuals in 1969–70 (although the wage equation does contain an *ad hoc* dummy for 1971–72). The lack of a positive residual for 1969 is not surprising, since the equation is based on annual data, and the 1969 annual total was only marginally affected by the wage explosion. The absence of a residual in 1970 does surprise us, in view of the absence from the explanatory variables ($U$, $\Delta U$, and $\Delta p^e$) of any variable that increased prior to the wage explosion. One possible explanation lies in the measure of the expected price increase, which was taken as "*U.K.* firms' expected change in the price of their [own] output estimated at the start of each year" (Data Appendix, p. 4). This is a quite illegitimate measure of price expectations in terms of the Manchester theory, since a high coefficient on this variable is consistent with the most rigorously

One alternative type of explanation points to international factors. It has already been noted that the inflation reimported in 1969 through the current-account surplus and reserve inflows was offset by domestic deflationary fiscal-monetary policies, but there remains the possibility that the improved competitiveness resulting from devaluation served to undermine the resistance to wage claims. Nordhaus, after all, in his comparative study of wage inflation in seven countries, found that the scope for wage increases provided by foreign competition provided the best single explanation of British wage inflation. Comparison of columns 1 and 10 does not suggest that this is an entirely adequate explanation, and there are also three particular problems in placing primary emphasis on this factor. First, the increase in profit margins before the wage explosion does not appear to have been particularly large, as this theory would lead one to expect.[5] Second, the theory does not explain why the wage explosion was delayed as long as late 1969, some two years after the devaluation that is claimed to have precipitated it. Third, there is the problem as to why wage increases were maintained at this new higher plateau in 1971–72, after the competitive gains from devaluation had been eroded (Table 1). This is explicable in terms of inflationary expectations having been adjusted upwards to a new higher level by then as a result of the wage explosion (although the Carlson-Parkin estimates of price expectations shown in col. 3 of Table 3 do not provide strong support for this interpretation), but not in terms of a theory which asserts that the wage-determination process is dominated by a concern to maintain international competitiveness. Hence, while we do not doubt that relative international prices constituted a permissive factor in the initiation of the wage explosion, we find difficulty in placing primary emphasis on this factor.

The other alternative type of explanation points to what Laidler and Parkin term "sociological" theories, such as the frustration of wage earners at having their expectations of real income growth disappointed, or the increasing militancy of trade unions in pursuing a set of mutually inconsistent real income claims. Our one substantive departure from the Laidler-Parkin analysis concerns their rejection of such hypotheses (p. 764). Indeed, if the sociological label is intended to imply that such theories imply economically irrational behavior, we would quarrel with the label. If workers are utility maximizers who perceive that their personal economic welfare can be promoted by acting collectively in trade unions, and if the wage bargains sought by unions are such as to maximize their members' welfare, the wage rise sought will be such as to equate the marginal utility of higher income with the expected marginal disutility of greater unemployment and of longer strikes.[6] Since

cost-push theory of inflation, provided that firms are reasonably competent at forecasting the wage increase that they will be obliged to concede. In particular, if, at the start of 1970, firms were aware that a wage explosion had started, and expected to raise their prices in consequence, one would have an explanation of why the Gray-Parkin-Sumner wage equation has no positive residual, but the equation would in no sense explain the wage explosion.

[5] For example, the ratio of profits of companies and public corporations to income from employment increased only some 7 percent in 1969 as compared to 1967, and was still some 9 percent below the level of 1965 (when, admittedly, capacity utilization was dis-distinctly higher). Source: *NIESR Review*, Table 8.

[6] Thus we disagree with the propriety of the assumption of Parkin, Sumner, and Robert Ward that "... the wage level is set by a bargaining process between labor unions and employers' federations with the objective of achieving and maintaining a cleared market" (p. 7). A union which confined itself to achieving what the market would pay anyway would have no *raison d'être* except to counter monopsony, and would fail to maximize the utility of its members unless it placed an infinite welfare weight on continued employment of the marginal worker. It could in our view be argued more plausibly that the purpose of a well-conceived incomes policy is that of persuading agents to

the marginal utility of income may plausibly be expected to depend on dissatisfaction induced by comparisons with peer groups or expectations based on previous experience with a group's existing level of real income, since the power of trade unions may vary, and since changes in the moral climate may alter the disutility occasioned to an individual by participating in actions that hurt others (specifically, by striking), there would seem ample scope for "sociological" factors to influence the wage claims lodged by rational economic actors. Wage claims are not the same thing as wage awards, but any plausible theory of bargaining suggests that they will be positively associated, particularly in the presence of an accommodating monetary policy.

Despite the pioneering contribution of Albert G. Hines, attempts at econometric verification of such theories have not fared well up to the present. However, this may be because the variables which the theory posits as being important are difficult to measure. What can be said is that this type of explanation seems to offer a more plausible explanation of the wage explosion of late 1969 than does any other. Real incomes had been held down in the preceding two years—not so much because of the terms-of-trade effect of devaluation (col. 12), as because of the need to provide for the improvement in the current balance. Unions did become more militant, as reflected in a readiness to inflict direct hurt on the public in ways that had previously been considered unthinkable.[7] The improvement in the balance of payments, which became a clear fact in the summer of 1969, eased the permanent-crisis atmosphere of the preceding years. There were conspicuous and large salary increases awarded to top salary earners in the public sector. The Labor government, having lost its battle with the unions to reform industrial relations, tacitly allowed its incomes policy to become a dead letter rather than pick another fight with an election approaching. While this explanation is uncomfortably *ad hoc*, it is at least consistent with the facts, which is more than can be claimed for the more aesthetically pleasing theories.

In the absence of major change in monetary-fiscal policies or further shocks from abroad, we would have expected the wage explosion to produce a rise in unemployment to a level above the natural rate and, thereafter, a gradual deceleration of inflation back to the international norm. Persistence of the factors that produced the wage explosion would have resulted in a higher unemployment rate, rather than a permanently higher rate of inflation. Unemployment did indeed rise, peaking in the first quarter of 1972 at 3.8 percent, a level marginally above the latest Manchester estimate of the natural rate.[8] However, from 1970 on, monetary expansion broadly accommodated the increased rate of inflation (col. 6). In 1970–71 monetary growth was financed almost entirely by reserve inflows rather than domestic credit expansion (col. 7); this reflected the speculative flight from the dollar as much as the success of devaluation, and is the one period when it could conceivably be argued

accept settlements that clear the labor market rather than maximize monopoly rents. (By this test the only British incomes policy that might claim to have been well conceived was that of 1964–66—which is not to claim that this policy enjoyed any success in terms of restraining inflation.)

[7] This was first manifested in the dustmen's (garbage collectors') strike of October 1969. In subsequent years there were strikes by electricity supply workers, postmen, teachers, and doctors (in addition to major strikes by groups without a strike-free tradition, such as miners and railwaymen).

[8] The first estimate of the natural rate produced by the Manchester Inflation Programme was 1.7 percent (Parkin, Sumner, and Ward, p. 27), based on 1956–71 data. This has since been revised up to 1.8 percent for the period 1952–67 and 3.7 percent for 1968–74 (Gray, Parkin, and Sumner, p. 43), the break being explained by the more generous social benefits available to the unemployed in the later period.

that the United Kingdom imported inflation through monetary channels.[9] Such an argument is not very convincing, however, inasmuch as when unemployment approached the politically sensitive figure of one million, the government promptly embarked on an extremely expansionary fiscal policy which was financed in large part by monetary expansion. This set the stage for the second acceleration of wage inflation in 1974–75.

There is no difficulty whatsoever in accounting for this second wage explosion. Virtually every model of inflation, other than the naive Phillips curve estimated on the basis of predevaluation data, would have predicted a major acceleration of wage inflation in 1973–74. Monetary expansion was extremely rapid (col. 6), even after making allowance for the offsetting impact of import price increases on the real value of the money supply—which actually produced a noticeable effect in 1973–74 (col. 8). Although the current balance deteriorated in nominal terms, this was attributable to a deterioration in the terms of trade caused principally by the world commodity boom, while the current balance in real terms stabilized in 1973 and improved in 1974, thus doing nothing to ease the pressure of demand (col. 5). In total, demand pressure increased as desired, and unemployment fell rapidly to a minimum of 2.2 percent at the end of 1973. Meanwhile import prices shot up (col. 9–11) as a result of the world commodity boom, the oil price increase, and inflation in other industrialized countries, and also because of the depreciation of the pound following the adoption of a floating exchange rate in June 1972. The floating of the pound was inevitable, given the prior erosion of the gains in price competitiveness achieved by devaluation, and the commitment announced in the spring of 1972 to place "growth" (by which was meant a rapid expansion in demand) ahead of preservation of the Smithsonian parity. Finally, the adverse movement in the terms of trade created a need for a significant curtailment of real consumption (col. 12), thus providing ample grist for frustration theories. In view of this mass of reinforcing inflationary impulses, the fact that requires explanation is not that a second wage explosion occurred, but that it was delayed as long as 1974. The only convincing explanation of this delay is the obvious one: that the statutory and relatively rigorous incomes policy introduced in November 1972 actually did restrain wages, at least until the policy collapsed simultaneously with the government in the wake of the miners' strike in February 1974.

There can be no serious doubt that inflation was aggravated in the period 1973–74 as a result of the increase in the world prices of commodities and of oil (surely the clearest example of cost-push inflation in history). Not only did these cost increases produce a direct impact on the price level, which could not have been avoided by any feasible exchange rate appreciation, but they also exerted indirect effects, through

[9] It has sometimes been suggested (for example, by Robert Mundell, p. 281) that British institutional arrangements preclude reserve flows exerting an influence on the domestic money supply, since the operation of the Exchange Equalisation Account (*EEA*) ensures that all reserve changes will be fully sterilized. This is not quite true. With a net inflow of funds, the *EEA* pays sterling to the banks against foreign exchange, adding to both the cash reserves and the deposits of the banks. Then, as the authorities hold almost no cash, they issue a quantity of Treasury bills roughly equal to the inflow. If these bills are taken up by the banking system, then its liquid assets will increase and the banks can expand their lending, and the money supply is increased. If, however, the issue of bills is taken up outside the banking system, then the impact effect of the foreign currency inflow on the banks' deposit and liquid assets will be offset, and the reserve increase will be sterilized. The reason that in the past sterilization was usual is that the Bank of England has tended to resist the changes in the price of gilt-edged stock that occurred when major bank purchases or sales of gilts were induced by changes in the liquid assets of the banking system. Thus sterilization was a consequence of the policy of stabilizing interest rates, and not of the existence of the *EEA*.

frustrating real income growth (col. 12) and, perhaps, through strengthening the position of the miners in their confrontation with the government. Nevertheless, the claim that Britain imported the bulk of this second wage explosion is singularly unconvincing. First, there is no justification for arguing that increased world prices of manufactured imports contributed to British inflation: had other industrialized countries controlled their inflations better, the result would simply have been a larger depreciation of the pound. Second, demand pressures were excessive by late 1973, unless one is prepared to dismiss the evidence of an increase in the natural rate of unemployment. Third, the expansion of demand was so rapid as to have created strains even if the natural rate had been as low as in the early postwar years.

### III. British and World Inflation

Our examination of the development of British inflation up to 1975 therefore leads us to assign the major responsibility to internal factors. Prior to 1967 Britain was an exporter, not an importer, of inflation. The 1969–70 wage explosion was primarily an internal matter, and such impact as the external sector had (via either the frustration hypothesis or mitigated competitive pressures) was largely a reimportation of inflation in response to devaluation rather than a reflection of the comparatively modest inflation that had by then got under way in the rest of the world, which was still well below the British rate. By 1971–72 inflation had stabilized on a higher plateau, well above that prevailing elsewhere, but the government was unwilling to accept the higher level of unemployment needed to achieve a deceleration. The consequent restimulation of demand led to a downward float of the pound and a rapid build-up of demand; these internally generated pressures were indeed reinforced by imported inflation in the years 1973–74. By 1975 inflation was out of hand, and regarded as a sufficiently serious problem for the level of unemployment to be allowed to rise, and for a new statutory incomes policy to be adopted. Under these twin pressures, inflation has since abated.[10]

The conclusion that British inflation was predominantly internally generated naturally suggests the question as to why Britain pursued policies that led to her suffering the worst inflation in the Western world. On the whole we would not subscribe to the view that this was primarily attributable to economic illiteracy on the part of the authorities, despite the undoubted truth in the charge that they underestimated the long-run importance of monetary factors. Of greater importance, in our judgment, is the fact that price stability has simply not been regarded as an objective of paramount importance, in the way that full employment has been. For many years inflation was feared primarily because of its effect in undermining the balance of payments,[11] and not because of a fear of disruptive domestic consequences, social or economic. As long as it was believed that a little more employment could be bought at the cost of a slightly higher inflation rate, the general view was that the inflation was a price worth paying so long as the balance of payments permitted. And as the evidence accumulated that the cost was not in fact a slightly higher inflation rate but rather a permanently accelerating inflation rate, there was that reluctance to face un-

[10] The unemployment rate passed the latest Manchester estimate of the natural rate the month before the incomes policy was adopted, so no one will be obliged to modify their theory of inflation as a result of the latest deceleration.

[11] Indeed, even after sterling had been floating for three years, the popular explanation of the £6 a week incomes policy sent to every household (*Attack on Inflation: A Policy for Survival*) could still claim that a principal reason for curbing inflation was that "Britain's leading competitors in world markets now have the edge on us. Their prices are increasing at only half the rate of ours. And the gap is widening" (p. 4).

pleasant realities which has been one of the recurring failings of postwar British economic management. Whether inflation reaccelerates in the future is essentially a question of whether it is now sufficiently feared to create a willingness to pay the necessary price.

Our interpretation of British experience is not consistent with the view that world inflation originated entirely in the United States and was transmitted to all other countries through the pegged exchange rate system. Hence one must conclude that it was a coincidence that inflation accelerated at much the same time in both the United States and Britain. In our judgment the acceleration of inflation in France in 1968 was also indigenous, and was a further coincidence. To the presumably rhetorical question as to whether the global acceleration of inflation in the late 1960's can have been a coincidence, we would therefore answer that there was indeed an element of coincidence. On the other hand, we would also judge that the other countries of Northern Europe imported their inflations, through the various channels discussed in the British case in Section II. (We are not sufficiently familiar with the experiences of countries outside Northern Europe to offer a judgment as to whether there were any further coincidences, or whether the experience of all others can be accounted for through international linkages.)

Does this mean that Britain and France must share the blame with the United States for igniting an inflation of global proportions? To that question we would answer "no." While Britain and France generated their own inflations, the fact that they were not reserve centers (in any serious way, that is to say) meant that their capacity to export inflation to the rest of the world was strictly limited. After relatively modest deficits they had no option but to devalue and reimport the inflation that they had previously exported. There was no similar constraint on the United States and, hence, no similar limit to the harm that policy errors on her part could inflict on the rest of the world. It was to remedy this situation that other countries urged adoption of asset settlement in the negotiations in the Committee of Twenty that were intended to reform the international monetary system, and it was primarily because the United States declined to concede this point that the reform negotiations failed. In the event the rest of the world did indeed get its safeguard against a recurrence of global inflation, although in the form of floating exchange rates rather than asset settlement. Not only does this imply that the more stable countries will not be obliged to import inflation as they were in the early 1970's, but it also implies that countries that generate their own inflations will be deprived of an alibi.

## REFERENCES


**E. M. Bernstein**, "Flexible Exchange Rates and International Adjustment," in R. Hinshaw, ed., *The Economics of International Adjustment*, Baltimore 1971.

**J. A. Carlson and M. Parkin**, "Inflation Expectations," *Economica*, May 1975, *42*, 123–38.

**G. Edgren, K. O. Faxén, and C. E. Odhner**, "Wages, Growth and the Distribution of Income," *Swedish J. Econ.*, Sept. 1969, *71*, 133–60.

**H. Genberg**, *World Inflation and the Small Open Economy*, Stockholm 1975.

——— **and A. K. Swoboda**, "Causes and Origins of the Current Worldwide Inflation," paper presented to the I.E.A. Conference held in Saltsjöbaden, Aug.-Sept. 1975.

**M. R. Gray, M. Parkin, and M. T. Sumner**, "Inflation in the United Kingdom; Causes and Transmission Mechanisms," paper presented to the Conference on the Monetary Mechanism in Open Economies, Helsinki, Aug. 1975.

**A. G. Hines**, "Trade Unions and Wage Inflation in the United Kingdom, 1893–1961,"

*Rev. Econ. Stud.*, Oct. 1964, *31*, 221–52.

D. E. W. Laidler and J. M. Parkin, "Inflation—A Survey," *Econ. J.*, Dec. 1975, *85*, 741–809.

R. A. Mundell, *International Economics*, New York 1968.

W. D. Nordhaus, "The Worldwide Wage Explosion," *Brookings Papers*, Washington 1972, *2*, 431–64.

M. Parkin, M. Sumner, and R. Ward, "The Effects of Excess Demand, Generalized Expectations and Wage-Price Controls on Wage Inflation in the UK," paper presented to the Conference on Wage-Price Controls, Univ. Rochester, Nov. 1973.

Bank of England, *Quarterly Bulletin*, various issues.

———, *Statistical Abstract*, Vol. I, 1970, and Vol. II, 1975.

*Economic Trends*, June 1974, and annual supplement, Dec. 1975.

H. M. Government, *Attack on Inflation: A Policy for Survival*, 1975.

*International Financial Statistics*, various issues.

*Monthly Digest of Statistics*, various issues.

*Motor*, first issue in May each year, 1967–75.

National Institute of Economic and Social Research, *National Institute Economic Review (NIESR Review)*, various issues.

# The Welfare Costs of Nonoptimum Pricing and Investment Policies for Freeway Transportation

*By* MARVIN KRAUS, HERBERT MOHRING, AND THOMAS PINFOLD*

A "pure peak load phenomenon" can be defined to exist when (a) the demand for a service varies from hour to hour, day to day, or season to season—periods of time too short for capital stock to be varied so as to keep price continuously equal to long-run marginal cost—and (b) storage costs are so high that no inventories can be held, thus requiring output at any instant of time to equal consumption at that instant. It can be shown (see Section I) that the pricing and investment rules for optimum provision of peak load commodities are essentially the same as those for commodities for which demand shifts, when they occur, take place once and for all. If these rules for efficient resource allocation are violated—and in practice, they almost invariably are—welfare losses result.

The basic objective of this paper is to provide rough estimates of the welfare losses involved in providing freeway transportation, a commodity characterized by both a peak load problem and increasing returns to scale. Section I describes the model, the pricing constraints considered, and the loss measure employed. Sections II and III discuss the results of the pseudo-empirical analysis. The former describes the cost data used and their implications for the characteristics of optimum roads and trip prices in the absence of pricing constraints. The latter presents the demand relationships employed and estimates of the welfare losses resulting from the constraints. The Appendix accounts more fully for the selection of demand and cost data.

## I. The Basic Model[1]

Three commodities are produced in an economy: $X_1$ and $X_2$, one-mile freeway trips during peak and off-peak periods, respectively, and $X_3$, a Hicksian composite commodity representing everything else. The $I$ utility-maximizing consumers have identical tastes and enter identically into an egalitarian social welfare function. Each derives utility from his own consumption of these three goods and owns a stock of $\rho$ units of a general purpose resource which yields a flow of $r$ units (henceforth dollars) per year of resource services. A head tax of $h$ dollars is imposed on each consumer leaving him with a net income of $r-h$ dollars. Head taxes serve only to cover whatever deficits (or to redistribute what-

* Assistant professor of economics, Boston College; professor of economics, University of Minnesota; assistant professor of economics, Dalhousie University, respectively. The research reported was initiated during Mohring's Ford Foundation Faculty Research Fellowship and was subsequently supported by the National Science Foundation, grants by the U.S. Departments of Transportation and Housing and Urban Development to the University of Minnesota, and from the Canadian Transport Commission to the University of Toronto and York University during Mohring's and Kraus's visits there, and by large amounts of free computer time from the University of Minnesota. An earlier version of this paper was presented at the 1972 winter meetings of the Econometric Society in Toronto. Since this paper may be of less interest for its results than for the case study it provides on how *not* to do pseudo-empirical research, we will not embarrass the many who gave us advice and solace by acknowledging their help.

[1] It is described more fully and generally in Mohring (1970).

ever surpluses) may arise in the production of $X_1$ and $X_2$. Regardless of trip production rates, one dollar of resource services can be converted into one unit of $X_3$. Resource services are also used as variable and capital inputs in producing trips.

The total variable cost of producing one-mile trips during period $j$ at an *annual* rate of $X_j/\alpha_j$ is $X_j f(X_j/\alpha_j k) = C_j$; $j = 1, 2$. Here, $\alpha_1$ is the fraction of the year during which peak demand conditions are in effect, $\alpha_2 = 1 - \alpha_1$, and $k$ is freeway capacity, the maximum number of person-miles per year that can be provided on a one-mile expressway with a capital outlay of $K = d + ek$ dollars per year.[2] For all positive $X_j < \alpha_j k$, $f'(\cdot)$ and $f''(\cdot)$ are positive. Thus, with this formulation, increasing returns enter the picture only if $d > 0$ in the capacity cost function.[3]

In taking an automobile trip, a consumer provides all[4] of the variable inputs required—gasoline and other vehicle operating costs and, most important, his own travel time. In addition to these costs, $f(X_j/\alpha_j k)$ per person-mile in period $j$, he faces a toll of $T_j$ dollars per mile. The sum $P_j = f(X_j/\alpha_j k) + T_j$ is the price of a one-mile trip in period $j$.

In maximizing social welfare, society can control four variables: the identical head tax $h$ on each consumer; the peak and off-peak tolls $T_1$ and $T_2$ (or, equivalently, the prices, $P_1$ and $P_2$); and $K$, the annual outlay on freeway capacity. It is subject to the resource constraint $R = Ir = C_1 + C_2 + K + X_3$, and to one or more pricing constraints.

A variant of the peak load model is used to study spatial inefficiency in the pricing of freeway transportation. Interpret $X_1$ and $X_2$ as one-mile trips on freeways with high and low unit capacity costs, respectively. The total variable cost of providing $X_j$ one-mile trips per year on road $j$ is $C_j = X_j f(X_j/k_j)$. Providing $k_j$, the annual capacity of road $j$, requires an annual expenditure of $K_j = d_j + e_j k_j$. We refer to this model as the two-roads model.

Figures 1–3 respectively depict the efficient pricing and investment rules for peak load commodities and the nature of two of the constraints considered—the single toll and deficit constraints.[5] Suppose that peak and off-peak periods each last six months. The specification of annual freeway capacity $k$ implies a six-month short-run marginal cost curve like $EFG$ in Figure 1. Also suppose that peak demand is independent of off-peak price and conversely. (These restrictive assumptions are dropped in the subsequent analysis.) Then peak and off-peak demand schedules can be drawn as $D_1$ and $D_2$, respectively. Efficient utilization of the expressway can be shown to require equating the price of each trip to its short-run marginal cost, $GX_1$ and $FX_2$ in peak and off-peak periods, respectively. With marginal cost

[2] Implicit in this formulation is the assumption that only one technology—one production function—is available to provide trips. This assumption would be harmless in the absence of the peak load phenomenon. With constant output rates, if different technologies can be employed in producing a commodity, the only one of relevance is that which at current factor prices involves the lowest unit cost. With peak loads, however, choice of the lowest cost technology can depend on the extent of seasonal demand variations. For example, the cost of highway capacity increases with increases in the extent to which crossroads involve overpasses rather than intersections at grade. With modest seasonal demand variations, full separation of access may be the cheapest way to provide trips between Here and There. With greater demand variations, cost minimization may dictate increasing the capacity provided per dollar spent by making some or all intersections at grade. The loss calculations described in Section III ignore such possibilities for adjusting variable and capacity cost functions. This being the case, they overstate to some degree the welfare losses involved in the pricing constraints considered.

[3] At least for highway travel, this characterization of costs seems reasonably realistic. See the Appendix discussion of cost function derivations.

[4] More accurately, "essentially all." We ignore those road maintenance costs that are directly attributable to road use.

[5] The pricing and investment rules for optimization subject to these constraints are developed formally in Mohring (1970).

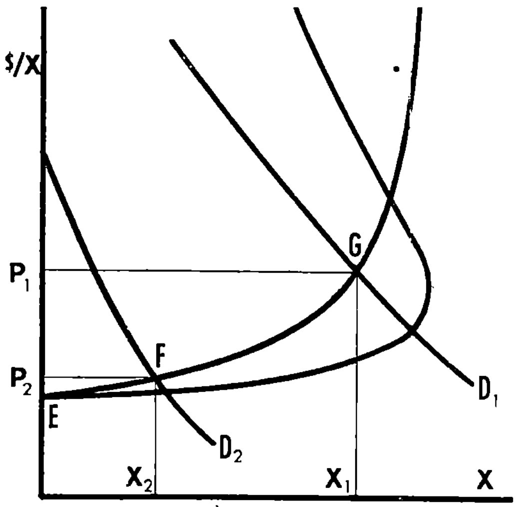



FIGURE 1

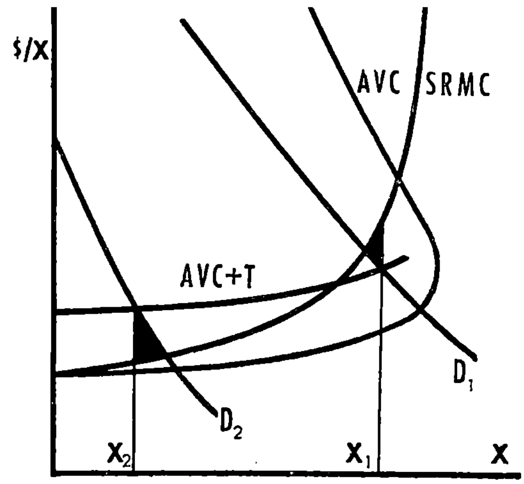



FIGURE 2

pricing, freeway capacity generates annual quasi rents of $EFP_2$ plus $EGP_1$. Efficiency also dictates that capacity be adjusted so as to equate the marginal cost $e$ of providing a unit of capital plant for one year with the annual quasi rents generated by that unit. In such a long-run equilibrium, total revenue just covers total costs if (and only if) the freeway is characterized by constant returns to scale, i.e., if $d=0$ in the capacity cost function.

The welfare costs of the following pricing and investment constraints are considered.

*Peak load—single toll:* The time required—again, probably the most costly input—for an expressway trip depends on the rate at which the expressway is used. However, a traveler is subject only to gasoline and other excise taxes which, per mile traveled, are nearly independent of traffic density. The constraint implied by this characteristic of excise taxes is that the price charged during a demand period must equal average variable cost during that period plus a toll $T$ that does not vary from period to period. If the toll yields a price of $AVC+T$ in Figure 2, too many trips will be taken during the peak period and too few during the off-peak. The resulting loss in real income equals (approximately)[6] the sum of the shaded areas in Figure 2.

*Two roads—single toll:* Efficiency would dictate higher traffic densities and hence higher tolls on high capital cost than on low capital cost freeways. However, gasoline and other excise taxes are invariant not only to the rate of travel on a given freeway but also to its cost. The same gasoline tax is imposed on trips over rural expressways costing perhaps $200 thousand a lane-mile to construct as on urban expressways costing $2 million a lane-mile. The constraint that the same toll per mile be charged on roads for which unit capacity costs differ yields a welfare loss similar to that depicted in Figure 2.

*Peak load—pure deficit:* While expressway construction appears to involve substantial scale economies (see the Appendix), user taxes are set to cover total costs. If demand is invariant over time and price equals average total rather than long-run marginal cost, a loss like $JKL$ in Figure 3 results. A similar conclusion applies when demand does vary from period

[6] The welfare loss measure employed in this study (see the discussion of Figure 4) does not involve the addition of consumers' and producers' surpluses. It does, however, yield similar numerical values.

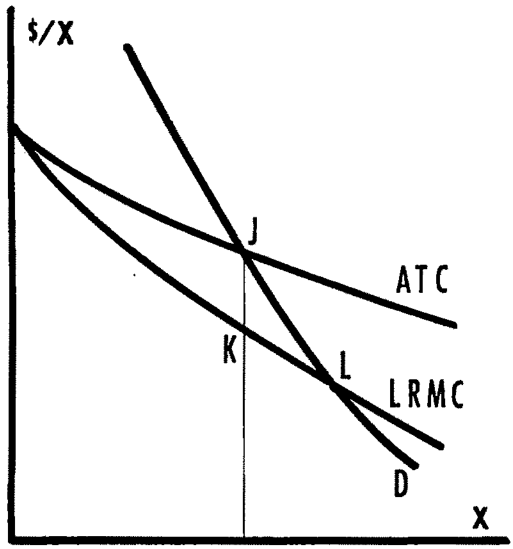



FIGURE 3

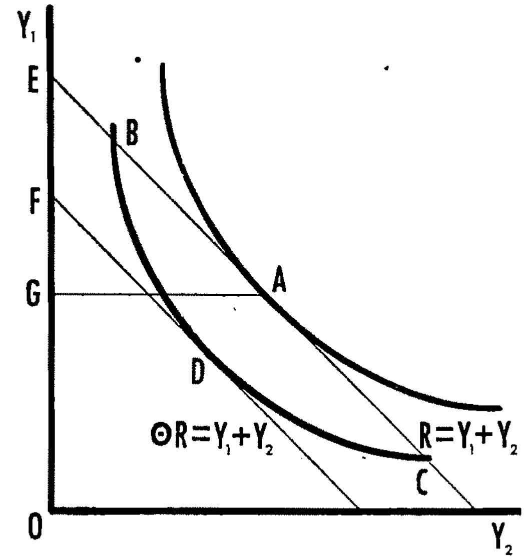



FIGURE 4

to period.

*Peak load—two roads—arbitrary capital input:* In treating the preceding constraints, it is assumed that freeway capacity and prices are optimized subject to the constraint in question. Actually, aggregate expenditures on highway capacity in the United States do not result from formal optimization procedures. Rather, they are largely determined by receipts from highway user taxes which in turn appear to be set arbitrarily. In terms of our model, this procedure implies specification of a single toll sufficient to cover the total cost of the highway(s) when capital inputs are arbitrarily specified.

*Peak load—single toll and deficit:* The peak load—single toll and deficit constraints are simultaneously imposed.

*Two roads—single toll and deficit:* The two roads—single toll constraint is combined with the constraint that total user taxes for the two roads just cover their combined capital costs.

The assumptions of identical tastes and a symmetric, egalitarian social welfare function permit analyzing the welfare costs of these constraints without reference to distributional considerations. For a one-consumer economy, Figure 4 depicts the type of index used to measure resource misallocation. With commodities 1 and 2 priced at marginal cost, a consumer would be in equilibrium at $A$. Imposing a constraint while at the same time adjusting his initial allocation of numeraire ($Y_1$, say) so as to keep him on the linear resource constraint, $R=Y_1+Y_2$, would force the consumer to an inefficient equilibrium such as $B$ or $C$. By simultaneously adjusting the price of good 2 and the consumer's initial endowment of $Y_1$, he could be moved to any other equilibrium on indifference curve $BDC$, $D$ in particular. At $D$, indifference curve $BDC$ is tangent to the budget constraint, $\theta R=Y_1+Y_2$. The loss $1-\theta$ (equals $FE/OE$) is an index similar to those suggested by Harold Hotelling, Arnold Harberger et al. of the proportionate decline in the consumer's real income caused by the pricing constraint. More precisely, $OF$ is the income that, at the prices which would be in effect under optimum conditions, would yield the same utility to the consumer as would $R$ units of resource services when prices are dis-

torted to produce an equilibrium at $B$ or $C$. The distance $FE$ can therefore be interpreted as the loss in his real income (valued at marginal cost prices) attributable to the distortion.[7]

The welfare loss measure employed in Section III is a three-dimensional counterpart of $FE/GE$, not of $1-\theta$ itself. That is, losses are measured as fractions of what the consumer's expenditures on the price-distorted commodities would have been without the distortions, not as fractions of his initial endowment of numeraire.

## II. The Characteristics of Optimum Freeways

Determining the welfare losses resulting from nonoptimum pricing and investment policies for expressways requires, *inter alia*, determining those road prices, operating characteristics, and the like which would optimize the unconstrained system. Specification of only variable and capital cost functions and equilibrium output rates suffices to determine these optimum road characteristics. The Appendix describes the reasoning underlying the specifications made. To summarize, average variable trip cost (in cents/person-mile) during period $j$ is[8] $3.67+V[1+(X_j/\alpha_j k)^{2.5}]$. Here, 3.67 cents is out-of-pocket vehicle operating cost for each of the 1.5 occupants of the average private passenger vehicle, and the bracketed expression is travel time in minutes/person-mile. The value of travel time $V$ is given alternative values equivalent to \$1, \$2.50, and \$4/hour.

Three groups of traffic flow conditions can usefully be distinguished in the typical urban area: morning and afternoon peaks; the period between about midnight and 6 A.M. when very few trips are taken (on the average 1970 weekday, 2.7 percent of the daily total in the Twin Cities area); and the remainder of the day. Rather than work with these three time periods, we have chosen to ignore completely early morning trips. The fraction of the remaining 18 hours of the day during which peak demand conditions prevail, $\alpha_1$, is assigned values of 1/6, 1/4, and 1/3. The values used for the ratio of peak to off-peak (i.e., other than peak and early morning) period trip consumption rates in the absence of pricing constraints, $(X_1/\alpha_1)/(X_2/\alpha_2)$, vary with $\alpha_1$. For $\alpha_1=1/6$, the ratios are 2, 2.5, and 3. For $\alpha_1=1/4$, they are 1.6, 2, and 2.4, while for $\alpha_1=1/3$, they are 1.4, 1.75, and 2.1. These ratios roughly reflect conditions prevailing in the Minneapolis–St. Paul metropolitan area.

If the annual cost of a one-mile freeway with a capacity of $k$ person-miles per year is $d+ek$, marginal cost tolls for a freeway of optimum size would generate annual revenues just sufficient to cover $ek$. An annual capital cost subsidy of $d$ per mile of freeway would therefore be required. If the individual studied is one of $I$ consumers each of whom travels $(X_1+X_2)/I$ miles/year, his "fair share" of this subsidy per mile traveled would be $g=d/(X_1+X_2)$. Three $(e, g)$ pairs are studied. The low pair (0.15¢, 0.5¢) reflects construction costs prevailing in the outer fringes of a large metropolitan area, while the high pair (1.5¢, 2.9¢) reflects costs adjacent to its central business district (*CBD*). The medium pair studied is (0.75¢, 2.1¢).

In the two-roads cases, average variable trip cost on road $j$ is $3.67+V[1+(X_j/k_j)^{2.5}]$ and its capacity cost is $d_j+e_jk_j$.

[7] If the resource constraint is linear and involves constant returns to scale *or* if it involves constant returns to scale and the utility function is homothetic, $\theta$ can also be interpreted as Gerard Debreu's "coefficient of resource utilization." That is, $1-\theta$ is the fraction by which the resources allocated to providing the consumer with goods 1 and 2 could be reduced while still yielding him the utility level associated with $B$ or $C$. In general, however, the Debreu loss measure would not be the same as that employed in this paper.

[8] Contrary to the discussion of Section I, computational considerations led us not to use a cost function which embodies the notion of "capacity" as a maximum flow rate. The function used does, however, reasonably approximate typical empirically derived cost-volume/capacity relationships over the range for which the marginal product of variable inputs is positive. See the Appendix.

TABLE 1—OPTIMUM PEAK LOAD ROAD CHARACTERISTICS

(Value of Travel Time=$2.50/hour)

| Assumed Traffic Pattern: | | Optimum Value of: | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Peak Period Duration | Travel Ratio[a] | Volume/Capacity[b] Peak | Volume/Capacity[b] Off-Peak | Toll/Person-Mile Peak | Toll/Person-Mile Off-Peak | Price/Person-Mile Peak | Price/Person-Mile Off-Peak | Subsidy/ Other Cost[b] | Subsidy/ Capital Cost[b] |
| | | | | Low Capital Costs | | | | | |
| 1/6 | 2.0 | 45 | 22 | 1.4¢ | 0.2¢ | 9.8¢ | 8.2¢ | 6 | 47 |
| | 3.0 | 48 | 16 | 1.7 | 0.1 | 10.2 | 8.0 | 6 | 42 |
| 1/4 | 1.6 | 39 | 24 | 1.0 | 0.3 | 9.3 | 8.3 | 6 | 48 |
| | 2.4 | 43 | 18 | 1.2 | 0.1 | 9.6 | 8.0 | 6 | 44 |
| 1/3 | 1.4 | 36 | 25 | 0.8 | 0.3 | 9.0 | 8.3 | 6 | 49 |
| | 2.1 | 39 | 18 | 1.0 | 0.2 | 9.2 | 8.1 | 6 | 46 |
| | | | | Medium Capital Costs | | | | | |
| 1/6 | 2.0 | 71 | 36 | 4.4 | 0.8 | 14.0 | 8.9 | 20 | 54 |
| | 3.0 | 76 | 26 | 5.3 | 0.3 | 15.3 | 8.3 | 19 | 49 |
| 1/4 | 1.6 | 62 | 38 | 3.1 | 1.0 | 12.2 | 9.2 | 21 | 55 |
| | 2.4 | 67 | 28 | 3.9 | 0.4 | 13.3 | 8.5 | 20 | 52 |
| 1/3 | 1.4 | 57 | 40 | 2.5 | 1.0 | 11.4 | 9.3 | 21 | 56 |
| | 2.1 | 62 | 29 | 3.2 | 0.5 | 12.3 | 8.5 | 20 | 53 |
| | | | | High Capital Costs | | | | | |
| 1/6 | 2.0 | 86 | 43 | 7.2 | 1.3 | 17.9 | 9.6 | 24 | 49 |
| | 3.0 | 93 | 31 | 8.7 | 0.6 | 20.0 | 8.6 | 22 | 44 |
| 1/4 | 1.6 | 75 | 47 | 5.1 | 1.6 | 14.9 | 10.0 | 25 | 51 |
| | 2.4 | 82 | 34 | 6.4 | 0.7 | 16.8 | 8.8 | 23 | 47 |
| 1/3 | 1.4 | 69 | 49 | 4.1 | 1.7 | 13.6 | 10.2 | 25 | 52 |
| | 2.1 | 76 | 36 | 5.2 | 0.8 | 15.2 | 8.9 | 24 | 49 |

[a] Ratio of peak period to off-peak period trip consumption rates.
[b] Shown in percent.

TABLE 2—OPTIMUM TWO-ROADS CHARACTERISTICS

| Travel Time Value | Optimum Value of: | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Volume/Capacity[b] Urban | Volume/Capacity[b] Rural | Toll/Person-Mile Urban | Toll/Person-Mile Rural | Price/Person-Mile Urban | Price/Person-Mile Rural | Subsidy/ Other Cost[a] | Subsidy/ Capital Cost[a] |
| | | | Low Capital Cost Difference | | | | | |
| $1 | 73 | 47 | 1.9¢ | 0.6¢ | 8.0¢ | 6.2¢ | 26 | 56 |
| $2.50 | 56 | 36 | 2.5 | 0.8 | 11.3 | 9.0 | 19 | 50 |
| $4 | 49 | 32 | 2.8 | 0.9 | 14.3 | 11.7 | 15 | 46 |
| | | | Medium Capital Cost Difference | | | | | |
| $1 | 82 | 43 | 2.5 | 0.5 | 8.9 | 6.1 | 31 | 56 |
| $2.50 | 63 | 33 | 3.3 | 0.7 | 12.4 | 8.8 | 22 | 49 |
| $4 | 55 | 29 | 3.7 | 0.8 | 15.6 | 11.4 | 17 | 46 |
| | | | High Capital Cost Difference | | | | | |
| $1 | 89 | 39 | 3.1 | 0.4 | 9.7 | 5.9 | 34 | 56 |
| $2.50 | 68 | 30 | 4.0 | 0.5 | 13.5 | 8.5 | 24 | 49 |
| $4 | 60 | 26 | 4.6 | 0.6 | 16.8 | 11.1 | 19 | 46 |

[a] On the assumption that 2.5 times as many trips are taken on urban as on rural roads; shown in percent.
[b] Shown in percent.

Three pairs of $e_j$ and subsidy/trip-mile (i.e., $g_j$) values are studied. One pair ($e_1=2.75¢$, $g_1=3.95¢$, $e_2=0.15¢$, $g_2=0.45¢$) is intended to reflect conditions which might prevail on a Manhattan expressway and a remote suburban freeway, respectively, while the second ($e_1=1.38¢$, $g_1=2.3¢$, $e_2=0.3¢$, $g_2=1.15¢$) and third ($e_1=2.06¢$, $g_1=3.15¢$, $e_2=0.23¢$, $g_2=0.8¢$) are meant to characterize two roads with more modest capacity cost differentials.

Tables 1 and 2 summarize optimum road characteristics for the peak load and two-roads models, respectively. To save space, optimum peak load road characteristics are tabulated only for the medium value of travel time, \$2.50/hour. The entries that would appear in tables based on the other travel time values studied can be determined fairly accurately from those shown in Table 1. If travel time is valued at \$1 an hour, the optimum peak period volume-capacity ratio is between 28–33 percent greater than that shown for the corresponding combination of parameter values. Optimum peak period tolls and prices with a \$1 travel time value are, respectively, 17–25 and 26–31 percent less than the corresponding entries in Table 1. For a \$4 travel time value, optimum volume/capacity ratios are 11–14 percent less while tolls and prices are, respectively, 11–15 and 21–30 percent greater than the corresponding entries in Table 1.

If Table 1 and the data from which it is drawn reasonably reflect real world conditions, tolls are substantially below their optimum peak period levels for most urban freeway travel while volume/capacity ratios are probably—reliable data on the subject are not available—substantially above the optimum levels. The calculations indicate that, if travel time is valued at \$2.50 an hour, a volume/capacity ratio in excess of 90 percent would be optimum only on high capital cost freeway segments and, even then, only when the peak period is brief and the ratio of peak to off-peak traffic flow rates is high. With travel time valued at \$1 an hour, volume/capacity ratios greater than 90 percent would be called for when capital costs are high and, with medium capital costs, when the peak period is of short duration. All other parameter value combinations studied yielded optimum volume/capacity ratios substantially below the near capacity levels which seem common in urban areas during rush hours.

The basic "toll" for highway travel in the United States—the sum of federal and state gasoline taxes—currently averages about 13 cents a gallon. The Bureau of Internal Revenue instructions for deducting state gasoline excises on federal income tax returns appear to be predicated on passenger vehicle gasoline use of 14 miles per gallon. These values suggest an average toll of 0.9–1.0 cent per vehicle mile or about 0.6 cents per person mile if the average private passenger vehicle occupancy rate is 1.5. Table 2 indicates this to be about equal to the efficient toll on an optimally designed rural road. However, with travel time valued at \$2.50 an hour, Table 1 indicates 0.6 cent per person mile to be only 7–15 percent of the efficient peak period toll on an optimally designed freeway near the *CBD* of a major urban area. And if travel time is valued at \$4 an hour, a toll 8–17 times that currently charged—the equivalent of a gasoline tax of \$1.00–2.20 per gallon—would be efficient during peak hours on a near-*CBD* freeway of optimum size.

## III. Welfare Losses

As Hotelling recognized, the welfare loss measure proposed in his famous 1938 article yields ambiguous results when applied to situations in which more than one market is distorted. That is, values of these measures depend on the order in which commodities subject to distortions (the two trip types in the case at hand)

are counted.[9] Provided that demand functions are integrable, the loss measure described in Section I does not suffer from this deficiency.

Unfortunately, the set of integrable systems of demand functions which both allow reasonable flexibility in specifying such attributes as own, cross, and income elasticities of demand *and* can be manipulated with a fair degree of ease appears to be empty.[10] The demand functions used in this study are those generated by the utility function

$$U = [\Sigma\epsilon_i(X_i + \beta_i)^\delta]^{1/\delta}$$

a *CES* function with a displaced origin.[11] For properly shaped indifference surfaces, $-\infty < \delta < 1$, and, without loss of generality, $\Sigma\epsilon_i = 1$. The implied demand functions are

$$(1) \quad X_j = -\beta_j + (w + \Sigma P_i\beta_i)/[(\epsilon_j/P_j)^\gamma Y], \quad j = 1, 2, 3$$

where $w = r - h$ is the consumer's income, $P$ denotes prices, $0 > \gamma = 1/(\delta - 1) > -\infty$, and $Y = \Sigma P_i(P_i/\epsilon_i)^\gamma$.

These demand functions afford considerably less flexibility in the specification of demand attributes than was desired. They do permit seven degrees of freedom in specifying these attributes. However, one of these serves only to fix the system's scale. It was hoped that the remaining degrees of freedom could be used to specify the values that would prevail in the absence of pricing constraints on the ratio of trip consumption rates $(X_1/\alpha_1)/(X_2/\alpha_2)$ and $X_1/X_2$ in peak load and two-roads models, respectively; the share of the consumer's income spent on trips $\rho$; and the own price and income elasticities of demand for the two types of trips, $E_{11}$, $E_{22}$, $E_{1w}$, and $E_{2w}$. However, for given values of the travel ratio and income share, it seems that there do not always exist combinations of utility function parameter values which yield desired combinations of these elasticities. Thus, we were unable to compute welfare losses under as wide a variety of demand conditions as we had hoped.[12]

[9] The same deficiency characterizes similar measures employed by Harberger and others. For a fuller discussion of this problem, see Mohring (1971).

[10] For example, suppose that constant elasticity demand functions apply to commodities one and two in a three-commodity world. With some effort, it can be shown that integrability requires goods one and two to have equal income elasticities of demand and rules out both gross substitutability and gross complementarity between them. Since it seems reasonable to suppose that a change in the price of a peak period trip would affect the quantity of off-peak trips consumed, this system was rejected. Similar analyses with other commonly used demand representations proved no more successful.

[11] This function bears the same relationship to the "linear" utility functions employed in studies by Lawrence Klein and Herman Rubin and by Richard Stone et al. as the *CES* function bears to the Cobb-Douglas.

[12] Using demand attributes to infer parameter values of the utility function which leads to equation (1) requires solving a non-linear equation system. We were unable to develop a reliable numerical procedure for doing so. We then turned to the following procedure for computing welfare losses under desired demand conditions: 1) do a series of preliminary runs to determine ranges of utility function parameter values that yield "reasonable" elasticity, etc., values; 2) draw a random sample of utility function parameter values from independent uniform distributions over these ranges; 3) determine for each combination of parameter values drawn the demand elasticities, etc., that would prevail in the absence of pricing constraints and the welfare loss resulting from each constraint for all possible combinations of high, medium, and low values of cost function parameters; 4) with the welfare losses resulting from the constraints as dependent variables, fit quadratic forms in cost and demand attributes; 5) estimate welfare losses under desired demand and cost conditions by evaluating these equations. The equations revealed demand attributes to be highly collinear. As a result, reliable forecasts could not be obtained under the full set of demand conditions we had hoped to consider. At this point, we concluded that equation (1) is insufficiently flexible and resigned ourselves to limiting the scope of the analysis. It appears, then, that one engaged in this sort of work has two choices, neither very desirable. 1) Employ integrable demand functions. While unambiguous loss values will result, the limited flexibility afforded by the demand specification may seriously limit the scope of the study. 2) Dispense with integrability. Greater flexibility will result, but so will ambiguous loss values. Before we choose 1) again, we will seriously consider 2). It could turn out that the ambiguity is not that great—that different paths yield quite similar loss values. If so, the penalty of ambiguity could well be worth the benefits of greater flexibility.

TABLE 3—DEMAND AND COST PARAMETERS USED IN WELFARE LOSS MODEL

| | Low | Medium | High |
|---|---|---|---|
| Peak Period (High-Cost Road) | | | |
| Elasticities ($E_{11}$, $E_{1w}$)[a] | (0.26, 0.25) | (0.52, 0.50) | (1.00, 1.00) |
| Off-Peak (Low-Cost Road) | | | |
| Elasticities ($E_{22}$, $E_{2w}$)[a] | (0.52, 0.50) | (1.00, 1.00) | (1.88, 2.00) |
| Income Share ($\rho$) | 0.10 | 0.15 | 0.20 |
| Peak Period Duration ($\alpha_1$) | 1/6 | 1/4 | 1/3 |
| Travel Ratio: | | | |
| Peak Load: $\alpha_1=1/6$ | | 2.50 | |
| $\alpha_1=1/4$ | 1.60 | 2.00 | 2.40 |
| $\alpha_1=1/3$ | | 1.75 | |
| Two Roads | 1.35 | 2.50 | 5.00 |
| Capital Costs | | | |
| Peak Load ($e$, $g$) | (0.15, 0.50) | (0.75, 2.10) | (1.50, 2.90) |
| Two Roads ($e_1$, $g_1$, | (1.38, 2.30 | (2.06, 3.15 | (2.75, 3.95 |
| ($e_2$, $g_2$) | 0.30, 1.15) | 0.23, 0.80) | 0.15, 0.45) |
| Travel Time Value ($V$) | 1.00 | 2.50 | 4.00 |

[a] Although concrete evidence on the subject is scanty, the prevailing impression seems to be that peak period demand elasticities are substantially lower than off-peak elasticities. The assumption of a 1:2 ratio seems as plausible as any.

TABLE 4—PEAK LOAD WELFARE LOSSES[a]

(Shown in percent)

| Constraint | Deviation | Peak Period Elasticities | Off-Peak Elasticities | Income Share | Peak Duration | Travel Ratio | Capital Costs | Travel Time Value |
|---|---|---|---|---|---|---|---|---|
| Single Toll | Low | 0.38 | 0.51 | 0.67 | 1.23 | 0.34 | 0.09 | 0.80 |
| 0.66 | High | 1.04 | 0.81 | 0.67 | 0.43 | 0.90 | 1.35 | 0.53 |
| Pure Deficit | Low | 0.61 | 0.85 | 1.19 | 1.12 | 1.26 | 0.10 | 2.26 |
| 1.14 | High | 1.94 | 1.31 | 1.10 | 1.15 | 1.07 | 1.58 | 0.73 |
| Single Toll and | | | | | | | | |
| Deficit | Low | 3.21 | 1.32 | 3.34 | 4.30 | 2.69 | 0.30 | 5.65 |
| 3.20 | High | 3.25 | 10.87 | 3.09 | 2.68 | 3.48 | 5.61 | 2.15 |
| Low Arbitrary | | | | | | | | |
| Capital Input[b] | Low | 22.45 | 16.65 | 18.67 | 19.36 | 17.38 | 7.40 | 22.18 |
| 18.42 | High | 14.12 | 30.20 | 18.09 | 17.98 | 19.20 | 25.17 | 16.03 |
| High Arbitrary | | | | | | | | |
| Capital Input[c] | Low | 8.22 | 5.85 | 8.34 | 9.71 | 7.42 | 2.04 | 11.20 |
| 8.21 | High | 8.32 | [d] | 8.09 | 7.46 | 8.80 | 13.01 | 6.57 |

[a] Ratio of "real income change" loss measure to transportation expenditure under optimum conditions.
[b] Capacity equals 50 percent of the optimum value under the single toll and deficit constraint.
[c] Capacity equals 150 percent of the optimum value under the single toll and deficit constraint.
[d] The iterative procedure used to optimize the constrained system did not converge.

In the sensitivity analysis which follows, welfare losses are computed by first finding a combination of utility function parameter values which, under the specified cost conditions, yields the specified demand conditions.[13] The cost and demand parameter values studied are shown in Table 3.

For the peak load and two-roads models, respectively, Tables 4 and 5 show the welfare losses which result from optimizing the system subject to each pricing constraint when all system attributes assume their medium values and also when system at-

[13] The demand conditions are those which would prevail in the absence of pricing constraints.

TABLE 5—TWO-ROADS WELFARE LOSSES[a]
(Shown in percent)

| Constraint | Deviation | High Cost Road Elasticities | Low Cost Road Elasticities | Income Share | Travel Ratio | Capital Costs | Travel Time Value |
|---|---|---|---|---|---|---|---|
| Single Toll | Low | 0.29 | 0.26 | 0.41 | 0.49 | 0.17 | 0.50 |
| 0.42 | High | 0.52 | 0.59 | 0.42 | 0.29 | 0.69 | 0.34 |
| Single Toll and Deficit | Low | 2.02 | 1.22 | 2.38 | 2.43 | 1.47 | 3.88 |
| 2.28 | High | 2.90 | 4.54 | 2.22 | 2.00 | 3.13 | 1.60 |
| Low Arbitrary Capital Input[b] | Low | 26.38 | 20.94 | 22.41 | 20.30 | 18.76 | 25.38 |
| 22.17 | High | 17.18 | [d] | 21.82 | 23.36 | 24.82 | 19.76 |
| High Arbitrary Capital Input[c] | Low | 8.02 | 7.12 | 8.63 | 8.12 | 6.66 | 10.63 |
| 8.53 | High | 10.06 | [d] | 8.48 | 8.63 | 10.24 | 7.22 |

[a] Ratio of "real income change" loss measure to transportation expenditure under optimum conditions.
[b] Total expenditures on the variable component of capacity, $e_1k_1+e_2k_2$, restricted to 50 percent of the optimum value under the single toll-deficit constraint.
[c] Total expenditures on the variable component of capacity restricted to 150 percent of the optimum value under the single toll-deficit constraint.
[d] The iterative procedure used to optimize the constrained system did not converge.

tributes deviate from their medium values individually.[14] In these tables, losses associated with uniformly medium values are found in the column headed "Constraint." Thus, comparison of the uppermost entries in the "Constraint" and "Peak Period Elasticities" columns of Table 4 indicates that, if $E_{11}$ and $E_{1w}$ are lowered from their medium to low values while all other variables are held at their medium values, the single toll constraint welfare loss declines from 0.66 to 0.38 percent.

Three groups of system parameters—capital costs, demand elasticities, and the value of travel time—are by far the most important determinants of welfare losses. An increase in either capital costs or off-peak or low-cost road demand elasticities leads to increases in all loss measures. An increase in the travel time value is associated with increases in *absolute* levels of resource waste, but decreases in the proportionate measures which are tabulated. The effects on losses of the income share, peak duration, and travel ratio parameters are, for the most part, modest and vary in direction from model to model.

Except when capital costs are low or the peak duration is short, the peak load single toll constraint appears to be substantially less consequential than the pure deficit constraint. Excluding from consideration only the low entries in Table 4 for these two parameters, the pure deficit constraint loss ranges between 1.2–3.7 times the corresponding single toll constraint loss for any given combination of

[14] For some parameter combinations, the first-order conditions for optimization subject to some constraints either have no solutions or yield, in peak load models, off-peak trip prices at which the consumer would desire to sell rather than buy them. This is particularly true of models involving deficit limitations when both capital costs and elasticities are high. This result seems attributable to the nature of the demand curves generated by the utility function employed. With this function, demand curves involving high elasticities at any one set of price and income levels are generally rather flat. More precisely, they generally intersect the price axis at a level not much greater than that through which the curve is made to pass in the initial specification of utility function parameter values. Under such circumstances, modest price increases can lead to the equivalent of supply and demand curves intersecting at negative or imaginary output levels.

parameter values.

There is some reason to believe (see Mohring 1965) that urban expressway capacity is substantially below the optimum level in the United States. If this belief is in fact correct, the low arbitrary capital input constraint is the most realistic of those studied. Among the models tested, it yields by far the largest loss values. Actually, these values may well be biased downward. During the height of morning and afternoon peak periods, expressways in many urban areas appear to be operating in the equivalent of the backward bending portion of the average cost curves drawn in Figures 1 and 2. The losses resulting from operating in this range undoubtedly exceed those given by our model—a model in which, to repeat, the marginal product of variable inputs is assumed positive for all positive input levels.

The optimization of freeway capacities subject to the current procedure of financing investments entirely out of gasoline and similar user excises would put the single toll and deficit constraint into effect. Depending on the specific combination of parameter values involved, such a change would reduce losses to 4 to 36 percent of the levels associated with the low arbitrary capital input model.

With rare exceptions, past attempts to measure the welfare costs of various types of resource misallocation have yielded estimates which have seemed trivial, at least in percentage terms. Except for some of the arbitrary capital input model results, this study continues the tradition. For example, the single toll constraint entries in Tables 4 and 5 range between 0.09–1.35 percent. This being the case, it is worth attaching dollar magnitudes to these percentage losses in partial answer to the almost inevitable question, "Why bother?" The highway transportation undertaken by households and business firms in the United States accounts for considerably more than 10 percent of *GNP*, i.e., considerably more than $100 billion. While 0.09–1.35 percent of $100 billion *is* a small fraction of *GNP*, it is also $90–1,350 million, hardly a trivial amount. And this range may well have a downward bias. The value of consumer-supplied travel time does not enter into *GNP* as it is conventionally measured. However travel time is a substantial part of the denominator in the welfare cost-total transportation expenditure ratios shown in Tables 4 and 5. Furthermore, peak load and two-roads pricing constraints are simultaneously in effect in the real world. They are not alternative constraints as assumed in this study. Had these constraints been imposed simultaneously, estimated losses would undoubtedly have increased, perhaps substantially. Finally, the cost figures used in this study apply to expressway transportation. Both the levels of and the differences between optimum peak and off-peak tolls and prices on city streets are substantially greater than those shown in Tables 1 and 2 (see, for example, Mohring, 1964). Thus, the losses associated with this important part of road transportation are undoubtedly greater than those suggested by Table 4.

Would it be worthwhile to take the steps necessary to eliminate the welfare losses identified in this study? It appears that a definite answer awaits further research. Eliminating the arbitrary capacity expenditure part of the low arbitrary capital input constraint would seem most clearly in order. But whether that constraint is in fact applicable to the real world is open to dispute. Neither current highway user charges nor the cost data employed in this study take into account air pollution or the costs that traffic noise and smells impose on those who live and work near traffic arteries.

Eliminating either the pure deficit or the single toll constraint would entail

costs not taken into account in our analysis. As for the former, the welfare loss associated with medium values of all parameters for the peak load-pure deficit constraint model amounts to 1.14 percent of what the consumer would spend on trips under optimum conditions. From Table 1, optimum conditions would involve a subsidy equal to about 20 percent of this expenditure level and hence a welfare gain of about 5.7 cents for the average subsidy dollar. Under the same conditions but with high peak period elasticities an 8.1 cent average gain per subsidy dollar would result from complete absorption by the public of the deficit. The two-roads models yield similar benefit ratios for subsidies.

No political body has available a system of lump sum levies to finance its activities, of course. Existing tax sources themselves introduce distortions. To determine whether a net social gain would result from partial or complete public subsidy of the fixed component of highway capacity costs would require information not presently available on the welfare cost of these tax related distortions.

A similar consideration applies to the single toll constraint. A variety of methods has been proposed to vary highway user charges from place to place and time to time (see, for example, Ministry of Transport). All of these pricing systems would involve administrative and capital costs that could generally be expected to be larger for systems with greater sensitivity. That is, the costs of establishing such a pricing system would almost certainly depend on how close it would come to achieving complete equality between prices and marginal costs through space and time. To specify the fraction of the single toll constraint loss that it would be efficient to eliminate would require information not presently available on the relationship between the costs and the sensitivities of alternative toll systems.

## Appendix

### *The Sources of Cost Functions and Parameter Values*

The highway engineering literature reports a variety of attempts to fit relationships of the general form $c=g(X/k)$ and $t=f(X/k)$ to data on vehicle operating costs $c$, travel time t, and the volume/capacity ratio $X/k$, for various types of roads.[15] Vehicle operating costs appear to be virtually independent of traffic densities on freeways and rural roads. For a vehicle traveling at a constant speed of greater than about 35 mph, gasoline consumption and other operating costs per mile traveled increase with speed. Realized speeds, in turn, decrease with increases in traffic densities. At the same time, however, increased congestion results in an increase in the frequency with which drivers must accelerate and decelerate. According to the American Association of State Highway Officials (*AASHO*) "Red Book," pages 100–25, the costs of these maneuvers tend almost exactly to offset the cost savings associated with lower realized speeds. As a result, regardless of the actual density of traffic, a driver's vehicle operating costs are about those he would incur at the speed he would choose to travel on a road with a zero volume/capacity ratio.

L. L. Liston and R. W. Sherrer report that, in 1973, the sum (net of taxes) of the fuel, oil, tire, plus half of the maintenance and depreciation costs of operating a 1973 standard size automobile in Baltimore was 7.17 cents per vehicle-mile. The corresponding figure for freeway travel would be somewhat less. If operating costs on freeways are on the order of 5.5 cent per vehicle-mile, then adjusting by a vehicle occupancy rate of 1.5 gives 3.67 cents per person-mile.

While the specific functional forms differ, most of the relationships between travel time and volume/capacity ratios in the highway engineering literature agree in two important respects. First, they imply that the law of diminishing marginal productivity holds for the variable input, travel time, at all positive traffic levels. That is, the differ-

[15] For references and a more extended discussion of these studies, see Mohring (1964).

ence between marginal and average travel times per trip is positive at even the lowest traffic densities. Second, these relationships imply that a given stretch of highway does, indeed, have a "capacity" in the sense of being able to produce trips at a rate no greater than some specifiable maximum. If the rate at which travelers attempt to make trips exceeds this level, travel time per trip continues to increase. However, the number of trips completed per time period actually falls; the marginal product of travel time is negative.

The simplest travel time-volume/capacity relationship which approximates those in the highway literature with reasonable accuracy is $X/k=4/\mathrm{t}-4/\mathrm{t}^2$ with t expressed in minutes per trip-mile. This expression underlies Figures 1 and 2. Working with a relationship that allows for the possibility of a negative marginal product for travel time presented computational difficulties that we preferred to avoid. We therefore settled on a relationship suggested by Robert Solow and William Vickrey (p. 436), $\mathrm{t}=a+b(X/k)^d$. For $a=b=1$ and $d=2.5$, this relationship yields travel times per person-mile which closely approximate those from the relationship cited above over the range for which travel time has a positive marginal product. Both relationships, for example, imply speeds of 60, 51, and 30 mph, respectively, at volume/capacity ratios of 0, 0.5, and 1. The ratios of the marginal cost tolls implied by the relationship actually used to those associated with $X/k=4/\mathrm{t}-4/\mathrm{t}^2$ are, however, considerably greater than one for $X/k$ values between about 0.5 and 0.85 and, of course, considerably less than one for $X/k$ ratios in excess of 0.95.

The literature contains no consensus on the rate at which drivers are willing to trade money for reduced travel time. The value most commonly used by the Bureau of Public Roads and state highway departments in benefit-cost studies is $1.55/hour for an assumed 1.8 occupants of the average private passenger vehicle. The seeming precision of this number is misleading. It reflects the view of the Committee responsible for the original 1949 version of the Red Book that $1/hour was as reasonable a number as any that might have been chosen and the fact that both prices and incomes rose between 1949 and 1960, the date of the Red Book's most recent revision.

If drivers choose travel speeds so as to minimize the sum of vehicle operating and travel time costs, data from Robley Winfrey suggest that, c. 1968, desired speeds of 45, 60, and 75 mph imply travel time values of $0.45, $1.73, and $6.08/hour, respectively.[16] By analyzing the choices of a group of British civil servants between slow, cheap, and fast, more expensive means of commuting, Michael Beesley inferred that members of the lowest pay group studied valued their travel time at about 31 percent of their wage rates. This ratio increased with wage rates to a high of 50 percent for the highest pay group analyzed. Similar results were obtained by Thomas Lisco. That is, values of travel time increased as a fraction of income for those with 1966 salaries of less than $10,000 a year. The ratio equaled about 50 percent for annual incomes in excess of $10,000. For an American currently earning $12,000 a year, these ratios would imply travel time values on the order of $2–$3/hour. To take these disparate estimates into account, the implications of alternative travel time values of $1.00, $2.50, and $4/hour are explored in the analysis.

Some rough data developed by John R. Meyer et al., pages 200–11, can be interpreted as indicating that, around 1960, the annual cost at 6 percent interest of the capital invested in right-of-way and construction for one mile of an $N$ lane urban freeway was approximately $Z=\$11{,}200+\$2{,}500D+\$(7{,}200+300D)N$, where $D$, which appears to range between 10–300 in the United States, is the "net residential density" of the area through which the freeway passes, i.e., population per acre of land actually used for residential purposes. In addition, annual right-of-way costs were on the order of

[16] Denoting the value of an hour's travel time by $V$, the total cost of traveling a mile on an uncongested road can be written $c=g(S)+V/S$, where $S$ is the traveler's desired speed. For $c$ to be a minimum, $dc/dS=g'-V/S^2=0$, or $V=S^2g'$. The travel time values given in the text were determined by inserting the appropriate values of $S$ and $g'$ into this equation.

$0.005DZ$. Data privately supplied by the Federal Highway Administration (*FHWA*) indicate that maintenance costs averaged approximately $2,500 per lane-mile of expressway in 1972. Between 1960 and the end of 1973, the *FHWA*'s composite highway construction cost index increased by a factor of 2.1 while its maintenance and operation cost index increased by a factor of 1.1 between 1972 and the end of 1973.

That a substantial proportion of both construction and right-of-way costs are independent of the number of lanes provided reflects the fact that a freeway includes paved shoulders on both sides of each set of traffic lanes, a median strip between opposing traffic streams, and buffer strips between the freeway and adjacent land. As a result, only about 33 to 50 percent, respectively, of the land area used by four- and eight-lane freeways is devoted to freeway lanes themselves.

The capacity of a four- to eight-lane expressway is about 1,800 vehicles per lane-hour or 2,700 passengers per lane hour at an average vehicle occupancy rate of 1.5. Multiplying the Meyer et al. and *FHWA* annual cost data by the appropriate cost indices and dividing by 2,700 passengers per lane-hour times 18 (see below) hours per day times 365 days per year yields, for average and marginal costs per person-mile of capacity:

| Net Residential Density | Marginal Capacity Costs (*MCC*) | Average Capacity Costs (*ACC*) Number of Lanes 4 | 6 | 8 |
|---|---|---|---|---|
| 10 | 0.14¢ | 0.25¢ | 0.21¢ | 0.19¢ |
| 100 | 0.68 | 1.34 | 1.45 | 1.26 |
| 200 | 1.61 | 4.63 | 3.62 | 3.12 |
| 300 | 2.89 | 8.52 | 6.64 | 5.71 |

In the absence of pricing constraints, marginal cost tolls for an optimum highway would generate revenues just sufficient to cover the component of its capacity costs that varies with highway width. The fixed component would be supported by a public subsidy. If the average volume/capacity ratio on the highway is $h$, then $(ACC-MCC)/h$ would be the average subsidy per person-mile traveled.

The calculations summarized in Table 1 yield optimum volume/capacity ratios which, for the medium value of travel time, average 0.25 and 0.49 for marginal capacity costs of 0.15 and 1.5 cents, respectively. These marginal capacity cost values would roughly correspond to those prevailing respectively in the outer suburban fringe and in a near *CBD* area of a large urban region. Highways in such areas would normally be of four and eight lanes, respectively. The respective average subsidies per person-mile associated with these roads—$(ACC-MCC)/h$—would be about $0.13/0.25=0.52$ cents and $1.41/0.49=2.9$ cents. The combinations of marginal capacity cost and subsidy per person-mile actually used in the analysis are: low, $MCC=0.15$ cents and subsidy $=0.5$ cents; medium, 0.75 and 2.1 cents; high, 1.5 and 2.9 cents.

In the two-roads analysis, three sets of values are used for the marginal capacity cost and subsidy parameters. In the first, the *MCC* and subsidy values respectively equal 0.15 and 0.45 cents for the low cost road while *MCC* is 2.75 cents and subsidy 3.95 cents for the high cost road. In the third, least extreme case, $MCC=0.3$ cents and subsidy $=1.15$ cents for the low cost road while $MCC=1.38$ cents and subsidy $=2.30$ cents for the high cost road. In the intermediate case, the respective values are 0.23, 0.8, 2.06, and 3.15 cents.

During the average 1970 work day, about 4.9 million person trips were taken by bus or private passenger vehicle in the Minneapolis-St. Paul metropolitan area. Grouping these trips into those taken during the three hours most heavily used for travel, the second most heavily traveled, and so forth, yields:[17]

| Highest 3 Hours | Second | Third | Fourth |
|---|---|---|---|
| 23.8% | 21.3% | 17.7% | 14.7% |

| Fifth | Sixth | Seventh | Lowest 3 Hours |
|---|---|---|---|
| 12.4% | 7.4% | 2.3% | 0.4% |

In dividing this distribution between "peak" and "off-peak" travel, it seems most

[17] These ratios, it should be noted, reflect suboptimum prices at least during peak hours but also quite likely suboptimum levels of highway investment. The effects of these deficiencies offset each other somewhat in using actual travel ratios to infer those under optimum conditions. The data summarized above were privately supplied by personnel of the Twin Cities area Metropolitan Council.

reasonable to us simply to ignore the 6 early morning hours, 12–6 A.M., which account for only 2.7 percent of average daily traffic ($ADT$), and to regard the highest 6 and next 12 hours as peak and off-peak periods, respectively. If these flow rates apply seven days a week (weekend travel is only slightly lower and slightly less peaked than weekday travel in most urban areas), $\alpha_1$ would equal 1/3 and the relative flow ratio $(X_1/\alpha_1)/(X_2/\alpha_2)$, would equal 1.73. These are approximately the numbers used for the long peak duration–medium flow ratio combination.

The short duration–medium flow ratio combination was derived from the following considerations. Urban peak hour traffic flows typically exhibit a substantial directional imbalance. Traffic is considerably heavier toward than away from the central business district during the morning peak. The reverse is true during the afternoon peak. Suppose that peak period travel in the low flow direction and off-peak travel occur at the same average rate. Reversible center lanes are infrequently employed. This supposition therefore implies that any given expressway lane is subject to peak conditions only 3 rather than 6 hours a day (i.e., $\alpha_1=1/6$ rather than 1/3) and that the difference between peak and off-peak flow rates during this shorter period is twice that suggested by the preceding paragraph, i.e., the relative flow ratio is 2.46 rather than 1.73.

Low and high deviations from this 2.46 flow ratio are based on the alternative assumptions that peak period reverse flows occur at rates characteristic of the highest and lowest three off-peak hours (respectively 17.7 and 7.4 percent of $ADT$ per three hours) rather than the average flow rate of the entire off-peak period. These percentages respectively imply peak hour–peak direction flows of 27.4 and 37.7 percent of $ADT$ per three hours and relative flow rates of 1.95 and 3.16. Approximately the same percentage deviations from medium flow rates are employed for $\alpha_1=1/3$ and an arbitrarily chosen middle value, $\alpha_1=1/4$ with an assumed medium flow ratio of 2.

## REFERENCES


**M. Beesley**, "The Value of Time Spent in Traveling: Some New Evidence," *Economica*, May 1965, *32*, 174–85.

**G. Debreu**, "The Coefficient of Resource Utilization," *Econometrica*, July 1951, *19*, 273–92.

**A. C. Harberger**, "Three Basic Postulates for Applied Welfare Economics," *J. Econ. Lit.*, Sept. 1971, *9*, 785–97.

**H. Hotelling**, "The General Welfare in Relation to Problems of Taxation and of Railway and Utility Rates," *Econometrica*, July 1938, *6*, 242–69.

**L. R. Klein and H. Rubin**, "A Constant-Utility Index of the Cost of Living," *Rev. Econ. Stud.*, 1948, no. 2, *15*, 84–87.

**T. Lisco**, "The Value of Commuters' Travel Time, A Study in Urban Transportation," unpublished doctoral dissertation, Univ. Chicago 1967.

**L. L. Liston and R. W. Sherrer**, *Cost of Operating an Automobile*, U.S. Department of Transportation, Federal Highway Admin., Washington 1974.

**J. R. Meyer, J. F. Kain, and M. Wohl**, *The Urban Transportation Problem*, Cambridge, Mass. 1965.

**H. Mohring**, "Relation Between Optimum Congestion Tolls and Present Highway User Charges," *Highway Res. Record*, 1964, *47*, 1–14.

———, "Urban Highway Investments," in R. Dorfman, ed., *Measuring Benefits of Government Investments*, Washington 1965, 231–91.

———, "The Peak Load Problem with Increasing Returns and Pricing Constraints," *Amer. Econ. Rev.*, Sept. 1970, *60*, 693–705.

———, "Alternative Welfare Gain and Loss Measures," *Western Econ. J.*, Dec. 1971, *9*, 349–68.

**R. Solow and W. Vickrey**, "Land Use in a Long Narrow City," *J. Econ. Theory*, Dec. 1971, *3*, 430–47.

**R. Stone et al.**, *The Measurement of Consumers' Expenditure and Behavior in the United Kingdom, 1920–38*, Vol. I, Cambridge 1954.

**R. Winfrey**, *Economic Analysis for Highways*, Scranton 1969.

**American Association of State Highway Officials** (*AASHO*), Committee on Highway Planning and Design Policies, *Road User Benefit Analysis for Highway Improve-*

*ments*, Washington 1960.

U.K. Ministry of Transport, *Road Pricing, The Economic and Technical Possibilities*, London 1964.

U.S. Department of Transportation, Federal Highway Admin., *Highway Statistics: 1973*, Washington 1975.

# Age, Experience, and Wage Growth

By EDWARD LAZEAR*

During the past decade, much has been said about the role that on-the-job training plays in augmenting one's stock of human capital.[1] Up to this point, however, little has been done to distinguish the effect of on-the-job training from that of aging on the increase in human wealth. The reason rests primarily on the fact that it is difficult to observe or even define in some appropriate way the amount of on-the-job training that an individual possesses. In this paper, a method is developed by which one may compare the effects of work experience to those of aging per se. The difference is then attributed to on-the-job training.

The analysis deals with the relationship between an individual's wage growth pattern and his employment history. If, as the human capital framework suggests, individuals increase their wealth by investing in on-the-job training, one might expect workers who spend less time on the job during a given period of time to acquire less human capital. Thus, it is expected that the growth rate of wages will be related not simply to an individual's chronological age, but also to the amount of time spent on the job during the period under consideration. If this is in fact the case, then part of the opportunity cost of being unemployed consists of a nonmeasured human capital component of the wage rate. The total cost of unemployment is then the sum of foregone observed earnings plus the value of the unobserved payment of human capital. By calculating the amount of human capital investment foregone during periods of unemployment, an estimate of the importance of on-the-job training is obtained. The major findings are:

1. Young workers receive approximately one-third of their total employment compensation in the form of human capital. The effect of current work experience on wage growth is therefore substantial.

2. For individuals in the 14–24 year-old age group, aging is an important determinant of wage growth. Its importance, however, declines with age. While for a 19 year old, one year of aging implies about twice as much wage growth as does a year of experience, by the time the individual has reached 25.2 years, the effect of experience exceeds that of aging per se.

3. The finding that wage growth is inversely related to previous work experience casts some doubt on the validity of Yoram Ben-Porath's neutrality assumption.

4. As expected, an increase in formal schooling is associated with more rapid wage growth. It is also found that schooling increases wage growth by slightly more than does the equivalent amount of work experience.

5. Consistent with previous work, it is found that an increase in the number of hours worked implies more rapid wage growth.

* Assistant professor of economics, University of Chicago, and research associate, National Bureau of Economic Research. This research was supported by grants from the National Science Foundation and the Rockefeller Foundation. I wish to thank Charlie Brown, Zvi Griliches, Robert Michael, Jacob Mincer, Sherwin Rosen, and James Smith for comments on an earlier version of this paper. I also thank Janice Platt for excellent research assistance.

[1] See, for example, P. M. Blau and O. D. Duncan, Barry Chiswick, John Hause, James Heckman, L. A. Lillard, Robert Michael and the author, Jacob Mincer (1962, 1974), Mincer and Solomon Polachek, and Sherwin Rosen (1973).

6. It appears that, *ceteris parabus*, nonwhites enjoyed more rapid wage growth over the period in question than did whites. This may be the result of institutional changes or of unobserved quality differences at the time of hiring which work in favor of white workers.

7. Married workers have higher initial wage levels, but slower wage growth than do nonmarried ones. This is presumably a reflection of differences in the life cycle location not captured by the included human capital variables.

8. Union membership at the end of the period is a positive factor in the determination of wage growth.

## I. A Model

Suppose that wage growth over time takes the form of

$$(1)\qquad W_{69i} = AW_{66i}\exp\{\gamma_i t + u_i\} = AW_{66i}\exp\{3\gamma_i + u_i\}$$

where $W_{69i}$ = hourly wage rate in 1969 in cents for individual $i$

$W_{66i}$ = hourly wage rate in 1966 in cents for individual $i$

$\gamma_i$ = average annual growth rate of wages which varies across individuals

$A$ = a wage shift parameter unique to this three-year period, but invariant across individuals. "$A$" should be thought of as a nonmeasured cohort and business cycle effect

$u_i$ = the random error term where $u \sim N(0, \sigma^2 I)$

The growth rate $\gamma_i$ depends on a number of other parameters, the most fundamental of which relate to aging and the acquisition of human capital. Let us then start with

$$(2)\qquad \gamma_i = \alpha_0 + \alpha_1(S_{69i} - S_{66i}) + \alpha_2(OJT_{69i} - OJT_{66i})$$

where $S_{69i}$ = highest grade of schooling completed in 1969 by individual $i$

$S_{66i}$ = highest grade of schooling completed in 1966 by individual $i$

$OJT_{69i}$ = individual's stock of on-the-job training in 1969

$OJT_{66i}$ = individual's stock of on-the-job training in 1966

with $\alpha_0$, $\alpha_1$, and $\alpha_2$ all positive

Although the data to be used in this analysis are quite explicit with respect to job experience, it still remains impossible to directly measure with any confidence the amount of on-the-job training acquired over this three-year period. It is possible, however, to approximate the change in the stock of on-the-job training if it is assumed that

$$(3)\qquad OJT_{69i} - OJT_{66i} = \delta_1(E_{69i} - E_{66i}) + \delta_2 E_{66i} + \delta_3 S_{66i} + \delta_4 Age_i$$

where $E_{66i}$ = initial amount of job experience in years held by individual $i$[2]

$E_{69i}$ = amount of job experience in 1969 for individual $i$, equal to $E_{66i}$ plus the years worked between 1966 and 1969 where a year is defined[3] as (weeks worked/52)

[2] For individuals whose job in 1966 was obtained immediately after school, $E_{66i}$ is exact in years. For individuals whose job in 1966 was the same as the job worked in the final year of high school, $E_{66i}$ equals the number of years since school completion to the survey in 1966 plus .33 years. For all other individuals, $E_{66i}$ equals the number of years between school completion and the survey. Stratifying these groups yielded results which did not differ significantly from those obtained below.

[3] Since experience is measured in terms of years, and since it can be argued that the human capital production during any given year of experience varies with age, the amount of previous experience, and schooling, it might be suggested that equation (3) includes some type of interaction which segments the $(E_{69}-E_{66})$ variable. It was found that when (3) was specified to include multiplicative interaction terms, the resulting equation did not differ significantly from the one obtained when (3) is used as specified.

$Age_i$ = the individual's age in years in 1966.

The coefficient $\delta_1$ is expected to be positive since individuals who spend more time working are more likely to acquire on-the-job training (this essentially is the requirement that the cost of learning be a convex function of the learning rate).[4] The coefficient $\delta_2$ should be negative since it pays an individual who plans to invest in on-the-job training to do so during his first years on the job. As previous experience increases, incremental investment in on-the-job training for a given year of experience should fall.[5] The sign of $\delta_3$ depends upon the marginal complementarity or substitutability of formal schooling and on-the-job training. Finally, $\delta_4$ is negative since, given schooling and experience, older individuals are less likely to invest in human capital.

The term $(E_{69i}-E_{66i})$ may be rewritten as $(3-TN_i)$ where $TN_i$ is the total number of weeks during the three-year period in which the individual did not engage in work divided by 52. Upon substitution, we obtain

$$(4)\quad OJT_{69} - OJT_{66} = 3\delta_1 - \delta_1 TN_i + \delta_2 E_{66i} + \delta_3 S_{66i} + \delta_4\, Age_i$$

so that

$$(5)\quad \gamma_i = \alpha_0 + \alpha_1(S_{69i} - S_{66i}) + \alpha_2[\delta_1(3) - \delta_1 TN_i + \delta_2 E_{66i} + \delta_3 S_{66j} + \delta_4\, Age_i]$$

or

$$(6)\quad \gamma_i = \theta_0 + \theta_1 S_{66i} + \theta_2 E_{66i} + \theta_3(S_{69i} - S_{66i}) + \theta_4 TN_i + \theta_5\, Age_i$$

where

$$\theta_0 = \alpha_0 + \alpha_2\delta_1(3) > 0 \qquad \theta_1 = \alpha_2\delta_3 > 0$$
$$\theta_2 = \alpha_2\delta_2 < 0 \qquad \theta_3 = \alpha_1 > 0$$
$$\theta_4 = \alpha_2(-\delta_1) < 0 \qquad \theta_5 = \alpha_2\delta_4 < 0$$

Substituting (6) into (1) and taking the *log* of both sides yields

$$(7)\quad ln\, W_{69i} = ln\, A + ln\, W_{66i} + 3[\theta_0 + \theta_1 S_{66i} + \theta_2 E_{66i} + \theta_3(S_{69i} - S_{66i}) + \theta_4 TN_i + \theta_5\, Age_i] + u_i$$

or

$$(8)\quad ln\, W_{69i} - ln\, W_{66i} = \eta_0 + \eta_1 S_{66i} + \eta_2 E_{66i} + \eta_3(S_{69i} - S_{66i}) + \eta_4 TN_i + \eta_5\, Age_i + u_i$$

where

$$\eta_0 = ln\, A + 3\alpha_0 + 9\alpha_2\delta_1 > 0$$
$$\eta_1 = 3\alpha_2\delta_3 > 0$$
$$\eta_2 = 3\alpha_2\delta_2 < 0$$
$$\eta_3 = 3\alpha_1 > 0$$
$$\eta_4 = 3\alpha_2(-\delta_1) < 0$$
$$\eta_5 = 3\alpha_2\delta_4 < 0$$

The "effect of aging" may be defined as that wage growth which, in the absence of work experience or investment in formal schooling, occurs simply as the result of the passage of chronological time[6] independent of cohort and business cycle effects. Thus, define $\gamma'$ as the rate at which wages would grow if $(E_{69}-E_{66})$ and $(S_{69}-S_{66})$ were both set equal to zero. Then

$$(9)\quad \gamma' = \gamma - \eta_3(S_{69} - S_{66}) + 3\eta_4$$

since when $TN=3$, $(E_{69}-E_{66})=0$.

From (1), (8), and (9) one may write that

[4] See my 1975 paper, and Rosen (1973) for a more complete discussion of this issue.

[5] This is the usual result of an optimal investment plan in a life cycle context. Both Ben-Porath (1967) and Rosen (1972) derive this implication. Heckman, and Mincer and Polachek, on the other hand, show that under reasonable assumptions this result will not hold.

[6] This is essentially nonmarket acquisition of human capital which the individual acquires as he ages. With enough data on nonwork activities, one could presumably define in some precise way exactly what "aging" is.

$$(10)\quad \gamma' = (\eta_0 + \eta_1 S_{66} + \eta_2 E_{66} + 3\eta_4 + \eta_5\, Age)/3 - (ln\ A)/3$$

Equation (10) is not identified if we are unable to estimate $ln\ A$. However, since it is reasonable that $A \geq 1$, it must be the case that

$$(11)\quad \gamma' \leq (\eta_0 + \eta_1 S_{66} + \eta_2 E_{66} + 3\eta_4 + \eta_5\, Age)/3$$

Thus, an upper bound to the effect of aging may be obtained.

Up to this point, only human capital variables of the most traditional types have been included in the wage growth equation. However, there are reasons to expect that wage growth will depend on other factors as well. Employers may be willing to pay more per hour to workers who work longer hours because of, say, daily set up costs. In light of the work by C. M. Lindsay, Jacob Mincer and Solomon Polachek, and Donald Parsons, it is reasonable to suppose that wage growth will be a function of the change in the number of hours worked between 1966 and 1969. In addition, to the extent that military experience offers an alternative method of acquiring human capital, the change in the amount of military experience should be included. Also, since there are many reasons why blacks may have different incentives to invest in on-the-job training than whites, race may be a factor in determining wage growth. Similarly, one may expect that marital status and union membership may affect wage growth. These variables are added to (8) so that it becomes

$$(12)\quad ln\ W_{69i} - ln\ W_{66i} = \eta_0 + \eta_1 S_{66i} + \eta_2 E_{66i} + \eta_3(S_{693} - S_{66i}) + \eta_4 TN_i + \eta_5\, Age_i + \eta_6 CH_i + \eta_7 D + \eta_8 CM_i + \eta_9 M + \eta_{10} U$$

where $D$ is a dummy variable set equal to 1 for white individuals,

$M$ is set equal to 1 for individuals who were married during 1966,

$U$ is set equal to 1 for individuals who belonged to a union in 1969,

$CH$ is the "usual" number of hours worked in 1969 minus the usual number of hours worked in 1966,

$CM$ is the number of years of military experience in 1969 minus the number of years of military experience in 1966.

## II. Estimation

This equation can be easily estimated with longitudinal data obtainable from the National Longitudinal Survey (*NLS*). The data selected for this study were on young men, 14 to 24 years old. Since we are trying to estimate the effect of experience on wage growth, postulating that missed work time represents missed on-the-job training, we would like to examine a group of individuals who undertake substantial investment in human capital. Since men tend to invest in on-the-job training more than do women, and the young invest more than the old, the desired effects are most likely to be observed when looking at young men. The estimated effects will therefore tend to be stronger for the group in question than for the working population as a whole.[7]

The original sample has records on 5,225 individuals. This had to be reduced to 1,996 observations to meet the following criteria: First, it was necessary that individuals in the sample have wage rates reported in both 1966 and 1969. Second, individuals who reported that their wage rate was either less than 50¢ per hour or greater than $10 were dropped on the grounds that reported wages in those cases were unlikely to be correct. Finally, observations were dropped for which there was incomplete information on variables used in this analysis.

[7] Specifically, the magnitude of the *OJT* effects should be larger for this group. To the extent, however, that variance is reduced by truncating the sample, the precision of estimation may be reduced as the standard errors increase.

Equation (12) was estimated by ordinary least squares (*OLS*).[8] The results were:

$$(13)\quad \ln W_{69} - \ln W_{66} = .87948 \underset{(.00617)}{+ .00548\, S_{66}} \underset{(.00644)}{- .02014\, E_{66}} \underset{(.01202)}{+ .04895\,(S_{69} - S_{66})} \underset{(.01456)}{- .04784\, TN} \underset{(.00653)}{- .02266\, Age} \underset{(.00068)}{+ .00276\, CH} \underset{(.02377)}{- .05173\, D} \underset{(.11183)}{+ .05668\, CM} \underset{(.02738)}{- .11089\, M} \underset{(.02318)}{+ .16406\, U}$$

$R^2 = .156$; $SEE = .4367$; $F(10, 1985) = 36.8$

(The figures enclosed in parentheses are standard errors.)

The equation yields a number of interesting results. First, the coefficient on $TN$ is negative and significant. This is consistent with the theory. Individuals who spend less time in the work force acquire less human capital in the form of on-the-job training. Since this equation holds formal schooling constant, this coefficient is not biased by the substitution of formal schooling for on-the-job training during the nonworked period. The term reflects the net foregone investment in human capital associated with dropping out of the work force.[9] The coefficients are more easily interpreted when converted by the following computation: Taking the *antilog* of (12), we may write

$$(14)\quad W_{69} = W_{66} \exp \{\eta_0 + \eta_1 S_{66} + \ldots + \eta_{10} U\}$$

Differentiating (14) with respect to $TN$ gives

$$(15)\quad \frac{\partial W_{69}}{\partial TN} = W_{66} \exp \{\eta_0 + \eta_1 S_{66} + \ldots + \eta_{10} U\}\eta_4$$

so that

$$(16)\quad \frac{\partial W_{69}}{\partial TN} = -14.8$$

for a single white nonunion worker who was not in the military between 1966 and 1969. (All other variables assume their mean values.)

Equation (16) implies that being out of the work force for a period of one year between 1966 and 1969 will cost the individual 14.8¢ per hour reflected as lower earning power in 1969.That is, the value of the individual's stock of human capital will be 14.8¢ lower from 1969 onward than it would have been had he not missed the year. During the year, the individual earns his observed pecuniary salary plus the value of this human capital. The present value of the human capital component of wages is easily calculated. Since earning power is increased by 14.8¢ per hour by a year's investment, and since this increase is enjoyed on every hour worked (or spent investing) until retirement, the present

[8] By hypothesizing that the change in hours worked enters the wage change equation we introduce simultaneity bias. The wage growth equation writes wage growth as a function of the change in hours. The supply of labor, however, would relate hours worked to wages and thus change in hours worked to the change in wages. Because of this, *OLS* yields biased estimates of the effects. The reader may be somewhat reassured to learn that when the *CH* term was deleted from the equation, none of the remaining coefficients were significantly altered.

[9] It might be argued that this coefficient is a reflection of differences in ability rather than in human capital stock levels. An attempt was made to standardize for the component of ability not held constant by inclusion of schooling, previous experience, and age. The *NLS* has information on virtually every individual in the sample which gives their test scores on an examination which was designed to test their "knowledge of the world of work." If this can be considered a proxy for ability, inclusion of this variable into equation (12) should affect the *TN* coefficient if the ability argument were correct. Estimation of this equation yielded results which were virtually identical in all respects to equation (13). In particular, the coefficient on the ability proxy was insignificant.

value of the human capital component of earnings is

$$(17)\quad PV = (2000)(14.8)\int_{0}^{T-1969} e^{-rt}dt$$

where $T$ is the date of retirement and the individual works 2000 hours per year.[10] If $r=10$ percent, $(T-1969)=45$, then the human capital earned during the year previous to 1969 is worth \$2927.11. An individual with observed earnings in 1968 of, say, \$5,000 would in reality be receiving \$7927.11 since he earns \$2927.11 worth of human capital.[11]

When computing the cost of a given amount of unemployment, one should add the value of foregone human capital to observed foregone earnings. The foregone earnings associated with being unemployed during 1968–69 can be estimated for this individual. If

$$(18)\quad \begin{aligned} W_{69} &= W_{66}\exp\{.42927\} \\ &= W_{66}\exp\{3\gamma\} \end{aligned}$$

then

$$(19)\quad \gamma = .143090$$

so that

$$(20)\quad \begin{aligned} W_{68} &= W_{66}\exp\{2(.143090)\} \\ &= \$2.67 \text{ per hour} \end{aligned}$$

Foregone earnings associated with missing 1968–69 amount to 2000 (\$2.67) or \$5340. Total foregone earnings are then \$5340+\$2927=\$8267 rather than the observed \$5340. The unobserved human capital component of earnings therefore is over one-half the size of observed earnings, or about one-third of total income. This is a very significant part of young men's salaries.

These results are important in that they reveal the existence of an experience effect. Work experience (or its complement) is related to wage growth independent of aging. Individuals who spend more time at work over the three-year period seem to experience more rapid wage growth which, it may be inferred, reflects on-the-job training. One may compare the effects of experience with that of aging per se. From (11) the upper bound of the aging effect was shown to be obtainable. After altering (11) to include the additional variables introduced in (12), we find that at the mean age of 19.334 years, the effect of aging on wage growth is at most

$$(21)\quad \begin{aligned} \gamma' &\leq (.27726)/3 \\ &\leq .09242 \end{aligned}$$

for a white unmarried nonunion worker. (All variables other than the dummies and $TN$ and $S_{69}-S_{66}$ assume mean values.) Thus, at age 19 the maximum effect of aging on wage growth is about twice that of experience (.09242/.04784=1.93).

One may obtain an approximation of $(ln\, A)/3$, which is the difference between the upper bound and true value of $\gamma'$, by the following procedure. In Table 1, regressions 1 and 2, wage level equations are estimated for 1966 and 1969. Since $A$ is a cohort effect which is not captured by variables included in (12), define

$$(22)\quad A = \frac{\hat{W}_{69}}{\hat{W}_{66}}$$

where $\hat{W}_{66}$ is the predicted value of wages allowing all right-hand variables in regression 1 to assume their mean values

[10] If foregone human capital affects nonmarket productivity as well as market productivity, this present value calculation will tend to understate the loss associated with unemployment. It is not likely that this understatement will be significant, however, since on-the-job training tends to be market specific in the type of human capital it provides.

[11] The individual might attempt to "catch up" with his uninterrupted age-earnings profile. However, if Ben-Porath's assumption of neutrality were correct, there would be no reason to engage in any type of catch-up behavior. The marginal cost of human capital would be independent of whether 1968–1969 were worked or not. As long as the date of retirement is not significantly altered, the marginal return to a unit of human capital would be left unchanged as well so that there would be no reason to alter the optimal investment plan.

TABLE 1—SUPPLEMENTARY REGRESSIONS

| Variable | Regression 1 Dependent Variable= $lnW_{66}$ | Variable | Regression 2 Dependent Variable= $lnW_{69}$ |
|---|---|---|---|
| $S_{66}$ | .06060 | $S_{69}$ | .05319 |
| | (.00561) | | (.00367) |
| $E_{66}$ | .02646 | $E_{69}$ | .01229 |
| | (.00582) | | (.00343) |
| *Age* in 1966 | .03724 | *Age* in 1969 | .02817 |
| | (.00571) | | (.00375) |
| Years of Military Experience as of 1966 | .03159 (.01102) | Years of Military Experience as of 1969 | .01149 (.00926) |
| Weekly Hours Worked in 1966 | .00433 (.00067) | Weekly Hours Worked in 1969 | −.00078 (.00081) |
| $D$ | .1876 | $D$ | .1542 |
| | (.02135) | | (.0180) |
| $U$ | .12154 | $U$ | .2866 |
| | (.02086) | | (.0177) |
| Married in 1966 | .18796 | Married in 1969 | .1163 |
| | (.01464) | | (.0179) |
| Constant | 3.324 | Constant | 4.0763 |
| $R^2$ | .454 | $R^2$ | .375 |
| *SEE* | .3972 | *SEE* | .3341 |

(except for dummies) and where $\hat{W}_{69}$ is obtained by plugging the 1966 means into the 1969 equation. This estimates what the typical young man in 1966 would earn if he were transplanted into 1969 and experienced the corresponding vintage effects. This procedure yielded an $\hat{A}=1.04$ so that $(ln\ A)/3$ equals .013. Thus, the upper bound of the aging effect does not differ substantially from our estimate of the true value of $\gamma'$.

Part of the aging effect may be attributed to a measurement bias. Since older individuals are less likely to invest in on-the-job training than are their younger counterparts, the observed wage understates the true wage (which includes compensation in the form of human capital) by a greater amount for younger individuals than it does for older ones. If so, a portion of the observed returns to aging would be illusory, resulting from this systematic bias in observed wages.

The last few paragraphs should not be taken to imply that experience is unimportant. It is clear that aging is important and understandably so for individuals in the 14–24 year-old age group. However, aging is parametric whereas experience is not. Experience has been shown to be important both in an absolute sense and relative to current wages. The fact that the effect of aging is so pronounced for the group in question is an interesting and useful result; it does not, however, negate the importance of the experience effect.

Age (not aging) enters negatively and significantly. Since the date of retirement is so far in the future, it is unreasonable to infer that the age coefficient reflects the difference in marginal returns to investment in human capital across individuals. Two explanations seem more plausible. First, the age component may reflect the influence of nonmarket time accruing previous to 1966. That is,

$$(23)\qquad Age \approx S_{66} + E_{66} + R_{66} + 6$$

where $R_{66}$ is defined as nonmarket time before 1966. The age coefficient is essen-

tially the effect of not attending school or working previous to 1966 on wage growth. However, its effect is an indirect one since most of the influence of $R_{66}$ should be captured by the initial wage $W_{66}$. The effect of $R_{66}$, therefore, relates to a quality component not held constant by the included variables. That is, a 24 year old who has the same amount of experience and schooling as an 18 year old (and thus has a higher value of $R$) obviously uses his time differently. If past work history is correlated with future labor force behavior, the older individual has a lower probability of being employed throughout the full year than does the younger one. The age coefficient then picks up the desire to spend a smaller proportion of each working year on the job and the consequent reduced investment in on-the-job training.

Alternatively, one may interpret the negative coefficient as an interaction term which reduces the effect of *aging* on wage growth as the individual ages. That is, the rate of wage growth which cannot be attributed to work experience or formal schooling ($\gamma'$) declines with age as shown above (although not necessarily relative to that which can be explained by experience and schooling). This is consistent with casual notions of physical and emotional maturation. The difference between a 15 and 16 year old may be considerably more pronounced than that between a 23 and 24 year old. The following calculation reveals the magnitude of this difference. At age 16, the upper bound of ($\gamma'$) is equal to .11760, (with all other variables excepting dummies assuming their mean values). By age 24, the upper bound of ($\gamma'$) is reduced to .05716 and by 26, $\partial W_{69}/\partial(E_{69}-E_{66}) > \partial W_{69}/\partial\ Aging\ (=\gamma')$. The importance of job experience relative to aging grows as the individual matures and overtakes the upper bound of the aging effect by age 25. Thus, the effect of *aging* on wage growth depends on age, but they are not analytically the same. The age coefficient should not be interpreted as the aging effect.

In the same vein, one can distinguish the effect of experience on wage growth from that of previous experience on wage growth. The effect of the former as reflected by the $TN$ coefficient represents the amount by which wages increase with additional work experience in the current period. The latter relates to the rate at which individuals will acquire on-the-job training in the current period for each unit of current experience. It was anticipated that individuals with greater amounts of previous experience will invest in less on-the-job training in the current period since it (generally) pays to invest in larger amounts of training during the initial years of work. Thus, two individuals who differ only in their amounts of previous experience are expected to have different investment rates simply because previous investment is likely to be a substitute for current investment. Since

$$\frac{\partial W_{69}}{\partial E_{66}} = \left.\frac{\partial W_{69}}{\partial E_{66}}\right|_{Age} + \frac{\partial W_{69}}{\partial\ Age}\cdot\frac{\partial\ Age}{\partial E_{66}} \tag{24}$$

and since $\partial Age/\partial E_{66}$ may be taken to be generally equal to one (see equation (23)),

$$\frac{\partial W_{69}}{\partial E_{66}} = [W_{66}\exp\{\eta_0 + \ldots + \eta_{10}U\}] \cdot(\eta_2+\eta_5) < 0 \tag{25}$$

Thus, the effect of previous experience is as anticipated.

This also sheds light on the question of neutrality. If the rental price of human capital were constant over all units of human capital once age is given,[12] and if the marginal cost of human capital were not a function of previous investment (i.e., Ben-Porath's neutrality assumption),

[12] This amounts to assuming that retirement is a function of age alone and is independent of the amount of time spent working over the lifetime. In light of the work done by W. G. Bowen and T. A. Finegan which shows that labor force participation tends to vary positively with the level of education for older workers, it is unlikely that the assumption is a valid one.

the effect of previous experience on the (absolute) change in wages would be zero. More simply, if neither the marginal cost nor marginal return to investment in human capital varies with previous experience, there is no reason for the experienced worker to behave any differently than the unexperienced one. The fact that experience does have a negative effect on $(ln\ W_{69} - ln\ W_{66})$ seems to imply that the neutrality assumption is invalid or that retirement (or more generally, time in the labor force) is a function of previous experience. This latter effect would be unlikely to account for much of the observed negative relationship since the retirement date for these individuals is expected to occur about forty years hence.[13]

There is an alternative explanation. Since the specification in (13) does not permit us to identify the effect of previous experience on on-the-job training from that of aging, it may be that the $E_{66}$ variable is a term which reflects the interaction between previous experience and current aging. If previous experience were in some sense substitutable for aging in the production of human capital, the negative relationship would be consistent.

The coefficient on incremental schooling is positive and significant. As has been argued elsewhere,[14] this should not be interpreted as a rate of return to education. It is, however, interesting to ask what the net effect is of dropping out of the labor force to attend school. The results in (13) permit this calculation. Since

$$(26) \quad TN = .75(S_{69} - S_{66}) + TO$$

where $TO$ is nonschool time not worked and the school year is assumed to be .75 years long. We may substitute so that

$$(27) \quad \left[\frac{\partial(ln\ W_{69} - ln\ W_{66})}{\partial(S_{69} - S_{66})}\right]^* = \frac{\partial(ln\ W_{69} - ln\ W_{66})}{\partial(S_{69} - S_{66})} + \frac{\partial(ln\ W_{69} - ln\ W_{66})}{\partial TN} \cdot \frac{\partial TN}{\partial(S_{69} - S_{66})}$$

where the expression on the left-hand side is the gross partial associated with an increase in schooling that occurs by dropping out of the labor force. Then

$$(28) \quad \left[\frac{(ln\ W_{69} - ln\ W_{66})}{(S_{69} - S_{66})}\right]^* = .04106 - (.04784)(.75) = \underset{(.0158)}{.00518}$$

where (.0158) is the standard error computed from the variance-covariance matrix.

Attending school is then slightly more productive in terms of wage growth than is on-the-job training. This is as it should be since the costs associated with the former are larger than those associated with the latter. In fact, the difference between the two may appear to be too small. This can be explained by the fact that the estimates relate to the average rather than marginal effects over the three-year period. If the ratio of average to marginal effects of schooling are less than the ratio of average to marginal effects of on-the-job training, then the estimate of the schooling effect overstates the marginal schooling effect by less than the estimate of the on-the-job training effect overstates the marginal on-the-job training effect. Under circumstances where the effect of schooling (on-the-job training) on the difference in the *log* of wages was a negative, convex function of previous schooling (on-the-job training), this would be the expected result since individuals in this group are at relatively high levels of formal schooling, but at low levels of on-the-job training.

Changes in the initial level of schooling have very small (and insignificant) effects on wage growth. To the extent that level

[13] See Brown for a more detailed discussion of the neutrality question.

[14] See Rosen (1973).

and increment are correlated, omission of the increment variable would lead to biases and to the incorrect conclusion that the rate of wage growth depended upon initial levels of schooling.[15]

As anticipated, an increase in the number of hours per week worked has a positive effect on wage growth. The partial effect is

$$(29)\quad \frac{\partial W_{69}}{\partial CH} = W_{66} \exp\{\eta_0 + \ldots + \eta_{10}U\}\ \eta_6 = .8521$$

so that increasing average weekly hours by 20 (i.e., moving from the typical part-time to the typical full-time job) increases wages in 1969 by about 17¢ per hour.

One rather interesting result is that, *ceteris paribus*, white workers experienced less rapid wage growth over this period than did black workers. This probably reflects the fact that 1966–69 witnessed rather significant changes in the institutional structure, one of the results of which was a narrowing of the black-white differential.[16] The signalling hypothesis provides another explanation.[17] It may be the case that blacks with equal schooling and job experience as whites may, by self-selection, be higher quality workers than are the corresponding whites. Employers who fail to discriminate in their reading of signals at the time of hiring will tend to pay the higher quality nonwhites the same wage as the lower quality white workers. Over time, however, employers learn about the nonwhite's relative advantages and the nonwhite worker's wages increase accordingly. (This would not be expected to persist over time, though, since employers who discriminated in favor of nonwhites with equal schooling and experience at the time of hiring would drive their less efficient competitors out of business.)

An equally bizarre result is that married workers experience less rapid wage growth than do unmarried ones. It is also the case, however, that their initial wage levels are higher (see Table 1, regressions 1 and 2). Since for this group (young men), holding age, schooling, and past experience constant, married men are in some sense farther along the life cycle path than are unmarried ones, the marriage dummy may reflect previous investment in human capital not captured by the included human capital stock variables. This would account for the higher initial wage level as well as for the lower wage growth since previous investment in human capital may result in less current investment as the outcome of nonneutrality.

Less surprising, perhaps, is the finding that individuals who belonged to a union in 1969 experienced more rapid wage growth. Part of this probably reflects higher wage levels for union workers since some of those who were union members in 1969 were nonmembers in 1966.[18]

Finally, the coefficient on incremental military experience is positive and of roughly the same magnitude as incremental schooling. The standard error here is large, probably because only 52 out of 1,996 individuals had a nonzero value for this variable. To the extent that the estimate is taken to be close to the true value, it appears as though one year in the military contributes as much to the stock of human capital as one year of formal schooling.

## III. Conclusion

This paper has estimated the human capital component of earnings and found it to be about one-third of total compensation for young workers. Although this im-

[15] David Wise examines the effect of schooling levels on wage growth. His analysis deals, however, with a very different sample and the findings are not easily compared.

[16] See Finis Welch for additional evidence on this point.

[17] See Kenneth Arrow and Michael Spence for a more complete discussion of the signalling hypothesis.

[18] Unfortunately, no information on union status in 1966 was available.

plies that human capital obtained through work experience is substantial at young ages, "aging" per se also accounts for a significant proportion of wage growth. This aging effect becomes relatively less important as the individual ages.

## REFERENCES


K. J. **Arrow,** "Models of Job Discrimination," in A. Pascal, ed., *Racial Discrimination in Economic Life,* Lexington 1972.

Y. **Ben-Porath,** "The Production of Human Capital and the Life Cycle of Earnings," *J. Polit. Econ.,* Aug. 1967, *75,* 352–65.

P. M. **Blau and O. D. Duncan,** *The American Occupational Structure,* New York 1967.

W. G. **Bowen and T. A. Finegan,** *The Economics of Labor Force Participation,* Princeton 1969.

C. **Brown,** "Young Men: Human Capital and Earnings," unpublished doctoral dissertation, Harvard Univ. 1974.

B. **Chiswick,** *The Distribution of Income,* New York 1974.

J. C. **Hause,** "The Covariance Structure of Earnings and the On-the-Job Training Hypothesis," Nat. Bur. Econ. Res. working pap. 25, Palo Alto 1973.

J. **Heckman,** "Estimates of a Human Capital Production Function Embedded in a Life-Cycle Model of Labor Supply," in N. E. Terleckyj, ed., *Household Production and Consumption,* Nat. Bur. Econ. Res. *Stud. in Income and Wealth,* vol. 40, New York 1974.

E. **Lazear,** "Human Wealth and Human Capital," Nat. Bur. Econ. Res. working pap. 97, 1975.

L. A. **Lillard,** "The Distribution of Earnings and Human Wealth in a Life-Cycle Context," Nat. Bur. Econ. working pap. 47, July 1974.

C. M. **Lindsay,** "Measuring Human Capital Returns," *J. Polit. Econ.,* Nov./Dec. 1971, *79,* 1195–1215.

R. T. **Michael and E. Lazear,** "On the Shadow Price of Children," paper presented at Econometric Society Meetings, New Orleans, Dec. 1971.

J. **Mincer,** *Schooling, Experience, and Earnings,* New York 1974.

———, "On-the-Job Training: Costs, Returns, and Some Implications," *J. Polit. Econ.,* Oct. 1962, suppl., *70,* S50–S79.

——— **and S. Polachek,** "Family Investments in Human Capital: Earnings of Women," *J. Polit. Econ.,* Mar./Apr. 1974, suppl., *82,* S76–S108.

D. **Parsons,** "The Cost of School Time, Foregone Earnings and Human Capital Formation," *J. Polit. Econ.,* Mar./Apr. 1974, *82,* 251–66.

S. **Rosen,** "Income Generating Functions and Capital Accumulation," disc. pap. 306, Harvard Univ. 1973.

———, "Learning and Experience in the Labor Market," *J. Human Resources,* Summer 1972, *7,* 326–42.

M. **Spence,** "Market Signalling: The Informational Structure of Job Markets and Related Phenomena," disc. pap., Public Policy Program, Harvard Univ., Feb. 1972.

F. **Welch,** "Black-White Differences in Returns to Schooling," *Amer. Econ. Rev.,* Dec. 1973, *63,* 893–907.

D. **Wise,** "Academic Achievement and Job Performance," *Amer. Econ. Rev.,* June 1975, *65,* 350–66.

"National Longitudinal Survey 1966–69," Ohio State Univ.

# Bond Share Yield Spreads Under Uncertain Inflation

By MYRON J. GORDON AND PAUL J. HALPERN*

It is generally accepted that individuals are concerned with the real and not the nominal values of income and wealth. Accordingly, Irving Fisher argued that the nominal rate of interest should be equal to the real rate plus the market's expected inflation rate; the real rate of interest is a function of production possibilities for the economy and individual's degree of time preference; in equilibrium, the real rate of interest is equal to the marginal productivity of capital.

However, Fisher's work did not recognize uncertainty and aversion to risk in the evaluation of investment opportunities. The work on capital asset pricing theory by William Sharpe, John Lintner, Jan Mossin (1966), and others explains the structure of asset yields and portfolio allocation decisions under uncertainty and aversion to risk, but their development of the theory has been based on nominal values of the variables. The implicit assumption is that the price level is stable or that the inflation rate is known with certainty. Either assumption is untenable in the current economic environment.

During the last few years a number of papers have appeared on the consequences of uncertain inflation for capital asset pricing theory. Marshall Sarnat has shown that linked bonds improve the efficiency of capital markets. Nahum Biger has examined the consequences for optimal portfolio selection. Richard Roll, and Andrew Chen and James Boness have extended the capital asset pricing model to recognize uncertain inflation, with a real riskless asset in one case and without it in the other.

This paper examines the consequences of inflation for the allocation of wealth between a monetary and a real asset and for the relation between the yields, both nominal and real, on the two types of assets. Section I defines the variables included in the analysis and states precisely the simplifying assumptions under which the problem is examined. Section II demonstrates that the allocation of an investor's wealth between a one-period government bond and a nonmonetary risky asset, for example, a share of stock, is a function of the uncertainty as to the inflation rate. Specifically, with the expected values of the real returns on the two assets given, the fraction of wealth invested in the bond decreases as the uncertainty of the inflation rate increases. Section III investigates what happens to the relation between the nominal and real yields on the two types of assets as the inflation rate uncertainty rises. We find that investors can only be persuaded to hold a given stock of monetary and nonmonetary (real) assets by a rise in the real return on the monetary asset in relation to the real return on the nonmonetary asset. In other words, the real return spread between the two assets decreases as the uncertainty as to the inflation rate rises.

Section IV examines the historical data on inflation over time and in different countries to obtain information on uncertainty as to the rate of inflation. These data are consistent with the hypothesis that uncertainty as to the inflation rate is an increasing function of the expected rate

* Professor and associate professor of finance, respectively, Faculty of Management Studies, University of Toronto.

of inflation. Hence, the mean realized inflation rate may be used as an index of the uncertainty as to the inflation rate. Finally, Section V examines some of the implications of the analysis.

Before proceeding to the analysis it may be noted that an interesting feature of it is the application of capital asset pricing theory to the macro problem of the structure of asset yields. Further work along the same lines may have considerable value for the analysis of macro problems of investment, income, and asset yields.

## I. Definitions

In the analysis we will assume that there are only two assets available to investors, a one-period government bond denominated in nominal terms with no default risk and a stock issued by a company.[1] In addition, we assume that the company has no debt in its capital structure, and its product pricing policies are flexible.

In the Fisher tradition we take the nominal return on a bond as the real rate of interest on the bond plus the rate of inflation, i.e., the percentage rate of change in the price level.

$$r_{Nb} = r_b + \Delta \tag{1}$$

where $\Delta \equiv$ the rate of change in the price level; $\Delta$ is a random variable with expected value $E(\Delta)$ and variance $\sigma^2(\Delta)$.

$r_{Nb} \equiv$ the nominal rate of interest on a bond.

$r_b \equiv$ the real rate of interest on a bond; $r_b$ is a random variable with expected value $E(r_b)$ and variance $\sigma^2(r_b)$.

Since the bond is a one-period bond denominated in nominal terms, its nominal rate of interest is not a random variable.

An analogous distinction between real and nominal yields on stocks can be drawn. The nominal yield on the stock, $r_{Ns}$, is equal to the real yield $r_s$ plus the rate of inflation. The expected value of the nominal return is

$$E(r_{Ns}) = E(r_s) + E(\Delta) \tag{2}$$

The variance of the nominal yield on stocks is

$$\sigma^2(r_{Ns}) = E\{[r_{Ns} - E(r_{Ns})]\}^2$$

But the realized nominal return on a stock is equal to the realized real return plus the actual rate of inflation. Therefore

$$\begin{aligned}\sigma^2(r_{Ns}) &= E[\{r_s + \Delta - E(r_s) - E(\Delta)\}]^2 \\ &= \sigma^2(r_s) + \sigma^2(\Delta) + 2\, cov\, (r_s, \Delta)\end{aligned}$$

where $\sigma^2(r_s)$ is the variance of the real rate of return on the stock.

We will assume here that $cov\, (r_s, \Delta) = 0$, which makes the stock a perfect inflation hedge.[2] Since we have assumed that the company has no debt, wealth transfers due to unanticipated inflation will not occur. The assumption further requires that the real return on the stock depends only on real risk factors which are independent of inflation. The consequence is that nominal earnings and stock price will vary with inflation, and

$$\sigma^2(r_{Ns}) = \sigma^2(r_s) + \sigma^2(\Delta) \tag{3}$$

One further concept is introduced here that will be used later. The assumption that $cov\, (r_s, \Delta) = 0$ logically implies that $cov\, (r_s, r_b) = 0$. Intuitively, variations in $r_b$ result only from variability in the inflation rate. However, fluctuations in $r_s$ are independent of the inflation rate. Therefore, $r_s$ is independent of $r_b$.

## II. Allocation of Wealth

Our purpose in this section is to establish the impact of inflation uncertainty on

[1] The stock can be thought of as a security representative of the market portfolio and a proxy for real or nonmonetary assets.

[2] This covariance is unlikely to be zero in fact. Withdrawing the assumption is an area for further research, but the qualitative nature of the conclusions reached is likely to hold except when the nominal return on the share is independent of the inflation rate.

an investor's wealth allocation between the stock and the bond. The investor is assumed to maximize expected utility and the form of his utility function is

$$u_i(r) = 1 - e^{-2\alpha_i r} \tag{4}$$

where $r$ is the real rate of return on the investor's wealth, $\alpha_i$ reflects the degree of risk aversion for investor $i$, and $\alpha_i > 0$. Bertram Schoner has shown that expected utility is given by the expression

$$E[u_i(r)] = 1 - e^{-2\alpha_i [E(r) - \alpha_i \sigma^2(r)]} \tag{5}$$

Therefore, maximizing expected utility is equivalent to maximizing

$$H(r) = E(r) - \alpha_i \sigma^2(r) \tag{6}$$

The choice of utility function was dictated by the following considerations. To keep the analysis mathematically tractable, it is necessary that expected utility be a function of the first two moments of the probability distribution of terminal wealth. To be realistic, the utility function should exhibit constant relative risk aversion. That is, the allocation of wealth between the bond and the share should be independent of the level of initial wealth. This utility function satisfies both conditions since $r = ((Y - W)/W)$, where $Y$ is random terminal wealth and $W$ is fixed initial wealth.[3]

Let $\lambda$ be the proportion of wealth invested in the stock and $(1-\lambda)$ be the proportion invested in the bond. The mean and covariance of the real return on the investor's wealth are

$$E(r_p) = \lambda E(r_s) + (1 - \lambda) E(r_b) \tag{7}$$

$$\sigma^2(r_p) = \lambda^2 \sigma^2(r_s) + (1 - \lambda)^2 \sigma^2(r_b) + 2\lambda(1 - \lambda)\, cov\,(r_s, r_b) \tag{8}$$

However, $cov\,(r_s, r_b)$, the covariance between the real returns on the bond and the share, is equal to zero.

Hence, to find the optimal proportions of wealth invested in each asset, we maximize expected utility or $H(r)$.

$$\max H(r) = \lambda E(r_s) + (1 - \lambda) E(r_b) - \alpha_i [\lambda^2 \sigma^2(r_s) + (1 - \lambda)^2 \sigma^2(r_b)] \tag{9}$$

The derivative of $H(r)$ with respect to $\lambda$ is

$$\frac{\partial H(r)}{\partial \lambda} = E(r_s) - E(r_b) - 2\alpha_i \lambda \sigma^2(r_s) + 2\alpha_i (1 - \lambda) \sigma^2(r_b) = 0 \tag{10}$$

and the optimal proportion of wealth invested in stock is

$$\lambda^* = \frac{E(r_s) - E(r_b) + 2\alpha_i \sigma^2(r_b)}{2\alpha_i [\sigma^2(r_s) + \sigma^2(r_b)]} \tag{11}$$

We are interested in investigating the impact of inflation risk on the allocation of wealth between stocks and bonds. The risk of inflation can be measured by the variance of the real rate of return on bonds. In the spirit of a partial equilibrium analysis, we assume that $E(r_b)$, $E(r_s)$, and $\sigma^2(r_s)$ are fixed and vary the uncertainty in

[3] Our use of equation (4) for a utility function has been questioned on the grounds that it exhibited constant absolute risk aversion and that the use of rate of return as the argument of the function is wrong. Precedents for these objections can be found in the literature. See Michael Jensen (p. 390) and Mossin (1973, pp. 39–40). However, we are at a loss to understand the grounds for these objections to equation (4). It can be written

$$u(Y) = 1 - e^{-\alpha((Y-W)/W)}$$

Taking the first and second derivatives of $u(Y)$ with respect to $Y$, we obtain for the Pratt measure of absolute risk aversion

$$Q = -\frac{u''(Y)}{u'(Y)} = \alpha/W$$

The Pratt measure of relative risk aversion is $\alpha$, which is independent of wealth. That is, the allocation of wealth between a risky and risk-free asset is independent of the investor's initial wealth. Initial and not terminal wealth is the relevant variable for the decision problem we are modeling and for which the Pratt measure is typically used. Finally it should be noted that the Pratt measure is a shortcut method for determining how an investor's investment in the risky asset varies with his wealth. Carrying out the numerical analysis with our utility function confirms the above description of it.

the inflation rate.[4] Since $\sigma^2(\Delta)=\sigma^2(r_b)$, we investigate whether $\partial\lambda^*/\partial\sigma^2(r_b)\gtreqless 0$ for small changes, where

$$(12)\quad \partial\lambda^*/\partial\sigma^2(r_b) = \frac{2\alpha_i\sigma^2(r_s) - E(r_s) + E(r_b)}{2\alpha_i[\sigma^2(r_s) + \sigma^2(r_b)]^2}$$

The impact on $\lambda^*$ can be seen most clearly if we consider different initial values of $\lambda^*$.

CASE 1: The investor is invested entirely in stock, i.e., $\lambda^*=1$. Then it is obvious that $\partial\lambda^*/\partial\sigma^2(r_b)=0$; the investor will not alter the proportion of wealth invested in the stock as inflation risk changes.

CASE 2: The investor holds both stocks and bonds in his portfolio, i.e., $\lambda^*<1$. Substituting this inequality into (11) we obtain $2\alpha_i\sigma^2(r_s)>E(r_s)-E(r_b)$. Substituting this inequality into (11) and noting that the denominator in (12) is positive, $\partial\lambda^*/\partial\sigma^2(r_b)>0$. Therefore, as inflation risk increases, the investor will increase the proportion of wealth invested in the stock.

CASE 3: The investor has shorted bonds to invest in the stock, i.e., $\lambda^*>1$. Using the same type of algebraic manipulations as Cases 1 and 2, it can be shown that $\partial\lambda^*/\partial\sigma^2(r_b)<0$. The investor will reduce the proportion of wealth invested in stocks by selling some stock and repaying the bonds he has created. As bonds become riskier, the debt he has undertaken becomes riskier and he reduces his involvement.

Therefore, in Cases 2 and 3, if an investor is in bonds as either a borrower or a lender, an increase in the variability of the inflation rate makes it attractive to him to borrow less if a borrower, or lend less if a lender.

### III. Impact on Yields of Both Stocks and Bonds

In the preceding analysis we assumed $E(r_b)$ and $E(r_s)$ were unchanged as $\sigma^2(r_b)$ increased. However, there will be a change in the yields on debt and equity as individual investors restructure their portfolios in response to a change in the variability of the inflation rate.

Before presenting a more rigorous analysis of the impact on real yields with a change in $\sigma^2(r_b)$, an intuitive explanation may be helpful. If individuals are unable to issue debt and the only debt outstanding is issued by the government, then $\lambda^*<1$ must hold and $\partial\lambda^*/\partial\sigma^2(r_b)>0$. Therefore, individual investors will move out of debt into stock. However, to persuade individuals to hold the exogenously given, unchanged quantity of debt and shares, the real yields on debt must rise and the real yield on the shares must fall.

The more realistic assumption that both individuals and government issue debt does not change the conclusion, since the net amount of debt in aggregate is the government debt. As $\sigma^2(r_b)$ increases, the real demand for bonds falls for investors with values of $\lambda^*<1$, and, for investors with $\lambda^*>1$, the real supply of bonds falls. The shift in the demand schedule will be larger in absolute amount than the shift in the supply schedule. Thus, the real rate on bonds will increase.

To investigate this more analytically, let $\bar{\lambda}$ be the allocation of market wealth which is a function of the weighted average of the individual investor risk-aversion parameters $\alpha_i$. Thus, from (11),

$$(13)\qquad \bar{\lambda} = \frac{E(r_s) - E(r_b) + 2\bar{\alpha}\sigma^2(r_b)}{2\bar{\alpha}[\sigma^2(r_s) + \sigma^2(r_b)]}$$

But $\bar{\lambda}$ is equal to the market value of the stock divided by the market value of stock plus debt since all assets must be held in equilibrium. The prices of the stock and the bond can be written $D/E(r_s)$ and $M/[1+E(r_b)]$, where $D$ is the expected value of the real dividend per share and $M$ is the expected value of the coupon plus principal on the bond. For simplicity, the

[4] The constraint that $E(r_b)$ remains fixed will be removed in Section III.

share's real dividend is not expected to grow, and the bond is assumed to be a one-period note. Thus,

$$(14) \quad \bar{\lambda} = \frac{n_s D/E(r_s)}{n_s D/E(r_s) + n_b M/[1 + E(r_b)]}$$

where $n_b$ and $n_s$ are the number of bonds and the number of shares outstanding.

If we assume further that the real coupon on the bond is equal to the real rate of interest, the bond is selling at par which is equal to unity. Therefore, (14) can be rewritten

$$(15) \qquad \bar{\lambda} = \frac{n_s D/E(r_s)}{n_s D/E(r_s) + n_b}$$

On the macro level, $n_s$, $n_b$, $D$, and $\bar{\lambda}$ are given. Equating the right-hand side of both equations (13) and (15) and solving we obtain

$$(16) \quad E(r_b) = E(r_s) + 2\bar{\alpha}\sigma^2(r_b) - 2\bar{\alpha}[\sigma^2(r_b) + \sigma^2(r_s)] \frac{n_s D}{n_s D + n_b E(r_s)}$$

To find the impact on yields as inflation risk increases, the derivative of equation (16) with respect to $\sigma^2(r_b)$ is taken.

$$(17) \quad \frac{\partial E(r_b)}{\partial \sigma^2(r_b)} = \frac{\partial E(r_s)}{\partial \sigma^2(r_b)} \left[1 + 2\bar{\alpha} \frac{n_s D}{[n_s D + n_b E(r_s)]^2} n_b[\sigma^2(r_b) + \sigma^2(r_s)]\right] + 2\bar{\alpha}\left[1 - \frac{n_s D}{n_s D + n_b E(r_s)}\right]$$

The quantities in the first and second set of brackets on the right-hand side of equation (17) are both positive.

To find the impact on the real return on bonds as inflation risk changes, *ceteris paribus*, set $\partial E(r_s)/\partial \sigma^2(r_b) = 0$. This results in $\partial E(r_b)/\partial \sigma^2(r_b) > 0$. Therefore, an increase in the uncertainty concerning the inflation rate will *increase* the real rate of return on bonds.

To determine the impact of inflation risk on the real return on stock, set $\partial E(r_b)/\partial \sigma^2(r_b) = 0$; this results in $\partial E(r_s)/\partial \sigma^2(r_b) < 0$.

Therefore, an increase in the uncertainty of the inflation rate will result in a reduction of the expected risk premium measured in terms of real rates of return.[5] By adding the expected rates of inflation to the real rates of return on stock and debt we can obtain the nominal yields.[6]

## IV. Relation Between Mean and Variance of Inflation Rate

This paper has demonstrated that the real rate of return on bonds is an increasing function of the variability of the inflation rate, and not the inflation rate itself. It is immaterial to the risk of a bondholder if inflation is 5 or 50 percent as long as the actual rate of inflation is equal to the expected rate of inflation. However, it is possible that the expected real yield on bonds is related to the expected inflation rate if there is a positive relationship between the expected inflation rate and the variability of inflation.

To investigate this possibility, values of the Consumer Price Index (*CPI*) were obtained for a sample of 33 countries for the period 1952 to 1972 inclusive.[7] The source for the data was *International Financial Statistics*. The countries were chosen so as

[5] Instead of assuming a one-period bond we could have analyzed the impact of changes in the inflation rate by using a consol. Although we did not perform the analysis it is clear that the impact would be in the same direction, but probably larger in magnitude because the impact of unanticipated inflation will be larger on a consol than on a one-period bond since the price of a consol is very sensitive to changes in the interest rate.

[6] It may be thought there is a logical contradiction in the statements that the nominal return on the bond is risk free and its nominal interest rate changes over time. This contradiction may be reconciled by assuming that all trading takes place only on the first day of each holding period, e.g., three months. Alternatively, with continuous trading the three-month return on a three-month bond is always certain, while the intra three-month return is uncertain. Investors nonetheless continuously look at the three-month return and treat the bond as risk free.

[7] A list of the countries used in the sample is presented in the Appendix.

to obtain a full spectrum of inflation rates. Annual percentage changes in the *CPI* were calculated for each year as well as the mean and standard deviation of the percentage change in the *CPI* for each country. The correlation between the mean percentage price change and the standard deviation over the 33 countries was .9081, and significant at the .001 level. The average annual realized rate of inflation for all countries over the sample period was .0848 and the average standard deviation was .0707.

Therefore, the differential between the expected real returns on bonds and stocks will narrow with increases in the expected rate of inflation. It is possible that in countries with hyperinflation, the expected real rate of return on bonds may exceed that on stocks. Since the nominal yields are obtained by adding the expected rate of inflation, similar conclusions on nominal yield differentials can be made.

## V. Implications

Two related conclusions follow from the previous analysis. One is that as uncertainty as to the inflation rate increases, investors move towards an all nonmonetary asset portfolio. The other conclusion is that insofar as the relative real quantities of bonds and nonmonetary assets fail to adjust to investor preferences, the spread between the yields on bonds and shares will narrow.

However, the analysis was carried out under a number of simplifying assumptions of which the following may be the most important. It was assumed that the uncertainty of the inflation rate is independent of its expected value. The previous section, however, established that the variability of inflation increases with its average rate. Hence, the portfolio adjustments investors desire as uncertainty increases are brought about more or less with no explicit portfolio revision by the declining real value of monetary wealth as the average inflation rate rises.

The other critical assumption was that the real return on nonmonetary assets is independent of the inflation rate. If the nominal returns on shares as well as bonds are independent of the inflation rate, there will be no desired movement between the two types of assets and no change in their relative yields. Under the more reasonable assumption that nonmonetary assets are at least a partial inflation hedge, the conclusions reached still hold, but with less force.

We may conclude with the question, what has happened to the spread between bond and share yields? Answering the question is made difficult by the problems of measuring share yields, but these problems are relatively modest for public utility shares. In the early 1960's, Aaa rated bond yields were about 4.5 percent, and electric utility share yields, measured by dividend yield plus expected dividend growth were about 8.5 percent; see Gordon (p. 99). Both authors have testified in recent utility rate of return cases, and we estimate current share yields for strong public utilities as being in the range of 12–13 percent. With Aaa rated bond yields around 9 percent, the absolute as well as the relative spread between nominal bond and share yields has fallen over the last decade.

## Appendix

*List of Countries*

| | | |
|---|---|---|
| Argentina | Finland | Norway |
| Australia | France | Peru |
| Austria | Germany | Portugal |
| Belgium | Greece | South Africa |
| Bolivia | Iceland | Spain |
| Brazil | Iran | Sweden |
| Canada | Israel | Switzerland |
| Chile | Italy | United Kingdom |
| Columbia | Japan | United States |
| Denmark | Mexico | Uruguay |
| Ecuador | Netherlands | Venezuela |

## REFERENCES


N. **Biger**, "The Assessment of Inflation and Portfolio Selection," *J. Finance*, May 1975, *30*, 451–67.

A. H. **Chen** and J. A. **Boness**, "Effects of Uncertain Inflation on the Investment and Financing Decision of a Firm," *J. Finance*, May 1975, *30*, 469–83.

I. **Fisher**, *The Theory of Interest*, New York 1967. (Reprinted 2d ed. 1930.)

M. J. **Gordon**, *The Cost of Capital to a Public Utility*, East Lansing 1974.

M. C. **Jensen**, "Capital Markets: Theory and Evidence," *Bell J. Econ.*, Autumn 1972, *3*, 357–98.

J. **Lintner**, "The Valuation of Risk Assets and the Selection of Risk Investments in Stock Portfolios and Capital Budgets," *Rev. Econ. Statist.*, Feb. 1965, *47*, 13–37.

J. **Mossin**, *Theory of Financial Markets*, Englewood Cliffs 1973.

———, "Equilibrium in a Capital Asset Market," *Econometrica*, Oct. 1966, *34*, 768–83.

R. **Roll**, "Assets Money and Commodity Price Inflation under Uncertainty," *J. Money, Credit, Banking*, Nov. 1973, *5*, 903–17.

M. **Sarnat**, "Purchasing Power Risk, Portfolio Analysis and the Case for Index-Linked Bonds," *J. Money, Credit, Banking*, Aug. 1973, *5*, 836–45.

B. **Schoner**, "Letter to the Editor," *Manage. Sci.*, Aug. 1967, *13*, B841–43.

W. F. **Sharpe**, "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk," *J. Finance*, Sept. 1964, *19*, 425–42.

International Monetary Fund, *Int. Financial Statist.*, various issues.

# The Two-Sector Open Economy and the Real Exchange Rate

By Michael Bruno*

This study arose from dissatisfaction with the multiplicity of models and separate theory compartments within which we usually teach and analyze the main aspects of trade, aggregate demand, capital inflow, and accumulation. It is proposed that in an open economy these can be tackled in one, rather simple, two-sector framework. It hopefully lends itself to a clearer exposition in some cases and possibly may also give new insights into some real phenomena. But first there are at least three mental hurdles to be overcome.

Traditional trade theory mainly assumed a world in which all goods (or the stylized "two goods") are in principle tradable and the main *relative* price involved is the international terms of trade. A third nontradable good sometimes appears as a relatively unimportant appendage to the standard two-tradable goods description. For most countries, however, international prices of imports and exports are given and a much more important relative price (or price vector) is the terms of trade between a composite foreign-exchange earning or foreign exchange using good and domestic nontradables. After all, the latter form over half of *GNP* even in the most open economies.

The importance of the tradable/nontradable distinction and the emphasis on the exchange rate as representing a real relative price have long been recognized in the development planning literature. But only in recent years have these aspects begun to play an important role in pure trade theory, particularly in the analysis of aggregate expenditure.[1] One elementary but essential difference between the two types of goods is that the market for nontradables must always balance *ex post*, that for tradables need not. The two-sector composition of excess demand is thus important for the understanding of the role of absorption and exchange-rate changes in short-run analysis of economic activity and inflation. This brings us to the next point.

Equilibrium in trade theory is usually identified with a balance on current account. There is extensive theoretical treatment of short-run deficits and surpluses, ways to correct imbalances, and the-once-and-for-all transfer problem, but there is much less systematic analysis of the prolonged (and often planned) *structural* imbalances that are a necessary by-product of substantial foreign borrowing. This can be viewed as an efficient way of relaxing

* Falk Institute and Hebrew University. The growth part of this paper is partly based on a more extensive unpublished M.I.T. discussion paper written in 1970. The static economy model (Section I) has been added since then. For financial support at the earlier stage I am indebted to the Ford Foundation (grants to M.I.T. and to the Israel Foundation Trustees) and to the Development Research Center of IBRD. I am greatly indebted to numerous colleagues for discussion and comments on earlier drafts; in particular I should mention Mordechai Fraenkel, T. N. Srinivasan, George Borts, and an anonymous referee.

[1] In a historical note on the survey by Anne Krueger, P. M. Oppenheimer traces the original introduction of nontraded goods to authors preceding I. F. Pearce's important contribution and going back to the late 1920's (Ohlin, Keynes). But I believe it fair to say that it never became central to the basic trade theory approach. Among the most recent studies with which the present one partly overlaps I should mention Rudiger Dornbusch, Eitan Berglas and Assaf Razin, and Elhanan Helpman.

period-by-period constraints on the market for tradables and replacing them by a single multiperiod foreign currency constraint.

The next point relates to the interface between pure growth theory and international trade. Once you allow trade, the consumption-investment goods distinction, so central to the closed-economy growth model, no longer retains its previous overriding importance. After all, exports or import substitutes are produced in order to earn (or save) badly needed foreign currency which can be put to many uses and can be borrowed and lent, so what does it matter whether one trades in nylon stockings or bulldozers?

Conforming to a common two-dimensional bias we start our analysis with a new, but essentially simple two-good descriptive model for an open economy. One good, measured in foreign exchange, is made to perform the multiple duty of juggling all tradable goods for consumption, physical accumulation, and storage of real financial debt. The other is a nontradable good that can be consumed or invested domestically. Given the production possibilities and consumption-demand patterns, the composition of output, price changes, and the balance of payments are essentially determined through absorption and exchange-rate policy decisions. Section I lays out the first, static-economy model and some of the policy implications for inflation and the deficit are analyzed. This is followed in Section II by an extension to a growth process, in which production possibilities and the real exchange rate are allowed to vary over time. Finally, the question of intertemporal optimality is considered (and a mathematical Appendix is given at the end). This has some practical implications for pricing and the rate of interest in public investment decisions.

## I. The Balance of Payments and Inflation in a Static Two-Sector Economy

Consider an open economy that uses and produces two basic types of goods: *nontradable goods* ($X_0$), which are only domestically produced, at the unit price $p_0$, and *tradable goods* ($X_1$), which can be purchased and sold abroad, at unit international price $P_1$ or domestic price $p_1$, where

$$p_1 = P_1 e \tag{1}$$

and $e$ is the nominal exchange rate.

The economy is here assumed to be small, and its international prices are given,[2] and also assumed to remain constant in relation to each other. This allows us to consider $X_1$ as a single composite good (rather than a vector of goods),[3] and $P_1$ as a given scalar which can be normalized so that $P_1=1$. In that case $X_1$ will also measure the potential supply of foreign exchange by the productive system.

The assumption of a single nontradable good (rather than a vector) can be defended on practical grounds if one remembers that nontradable goods (such as construction and services) are usually large users of domestic labor so that their prices are highly correlated with unit wage costs.[4] Both $X_i$ ($i=0$, 1) are measured in terms of value-added (total nominal $GDP=p_0X_0+p_1X_1$).[5] Any intermediate goods in the input-output sense are assumed to have been suitably netted out and added to $X_0$ and $X_1$, respectively, wherever appropriate.

At any moment of time (t) the supply is given of fixed reproducible factors (say nontradable and tradable assets $K_0$ and $K_1$, to be introduced in Section II) and of pri-

[2] This assumption can be relaxed.

[3] Transport costs are ignored here. Thus import substitutes and exports can formally be treated the same. Disaggregation is taken up in Section II.

[4] There may, of course, be some exceptions (e.g., electric power) but it still seems empirically reasonable for the bulk of the group consisting of services and construction, as long as productivity growth is fairly evenly spread within the group.

[5] $GNP$ can be derived by subtracting net interest payments on foreign debt (see below).

mary factors ($L$ being labor and/or land). Production possibilities in terms of $X_0$ and $X_1$ can be presented by the economy's transformation or production possibility curve ($PPC$) assumed to have a well-behaved convex shape:[6]

$$F(X_0, X_1; Z) = 0 \tag{2}$$

$Z$ is a shorthand notation for the vector of all factors that are fixed in the short run but may change in a long-run growth situation. The elements of the vector may include the momentarily given sectoral distribution of nonshiftable capital goods ($K_i$), the total supply of shiftable primary factors (such as $L$), and time (t), representing exogenous shifts such as "technical progress."[7] Underlying the aggregate $PPC$ (2) there must be short-run optimization of intersectoral allocation for at least one factor of production (say, $L$).[8]

The negative slope of the $PPC$, $\partial X_0/\partial X_1|_Z$ at any point measures the marginal rate of substitution in aggregate production between $X_0$ and $X_1$, which has the dimension of a real rate of exchange (domestic currency per unit of foreign currency) $q=p_1/p_0=e/p_0$.

Suppose now that the government can enforce a momentary rate $q$ in production at point $\tilde{M}$ on the $PPC$ (see $PP_1$ in Figure 1 and the line $BB_1$ touching the $PPC$ at point $\tilde{M}$). Total real $GDP(Y)$ can be read from the distance $OB$ in Figure 1 since we have (in units of the nontradable good):

$$Y = X_0 + qX_1 \tag{3}$$

(or $OB=OM+qOM_1$, since $q=MB/M\tilde{M}=OB/OB_1$).

For a given $Z$, $q$ uniquely determines the outputs $X_i$ ($i=0, 1$). For an increase in $q$, keeping $Z$ constant, we get the following price response (using subscripts for partial derivatives): $X_{0q}<0$, $X_{1q}>0$, $Y_q=X_1>0$ (we have $X_{0q}/X_{1q}=-q$).

Next consider household consumption demand. Denote total real household consumption by $C$ measured in nontradable units, and assume it is a portion $(1-s)$ of real $GDP$:

$$C = (1 - s)Y \tag{4}$$

$s$ is a measure of the private-sector contribution to the surplus, combining voluntary savings as well as government (direct) taxation. While government can be assumed to control overall $s$, the division of total $C$ between nontradables ($C_0$) and tradables ($C_1$) is based on consumer choice. Either we start from maximization of a utility function $U(C_0, C_1)$ subject to an expenditure budget constraint $C=C_0+qC_1$ or else by direct assumption, aggregate consumption functions can be written:

$$C_i = C_i(C, q) \qquad (i = 0, 1) \tag{5}$$

We have nonnegative marginal propensity to consume ($MPC$) (out of total consumption), i.e., $C_{ic}\geq 0$ ($i=0, 1$) if both goods are normal (and $C_{0c}+qC_{1c}=1$) and we also assume $C_{0q}\geq 0$, $C_{1q}\leq 0$ for the corresponding partial price responses.[9] Given $q$ (and thus $Y$) and given $s$ (and thus total $C$), the consumption basket ($C_0$, $C_1$) is determined uniquely. In Figure 1 this is represented at point $\tilde{N}(C_1, C_0)$ on the budget line $AA_1$, where $AB/OB=s$, $OA=C$. The projection of the consumption and production points on the respective axes give a measure of the residual commodity supplies (or surpluses) $S_i=X_i-C_i$, $NM$ and $N_1M_1$, respectively, for $i=0, 1$. For the response of these net supplies to changes in the policy parameters $s$ and $q$, we write:

$$S_i = X_i(q) - C_i[(1 - s)(X_0 + qX_1), q] \tag{6}$$

[6] Such a curve can be constructed empirically, e.g., from a detailed multisectoral linear programming system (see, for example, my 1967 paper). In that case $PPC$ will consist of a series of piecewise linear segments.

[7] It might also reflect shifts in world markets.

[8] The simplest model for which this can be derived is the standard two-sector production function model in which the sectoral $K$ are nonshiftable and $L$ is optimally allocated between the two sectors.

[9] Note that $C_0$ and $C_1$ need not be interpreted as final goods only, e.g., we can have total consumption ($C$) nontradable but using importables ($C_1$) in its production.

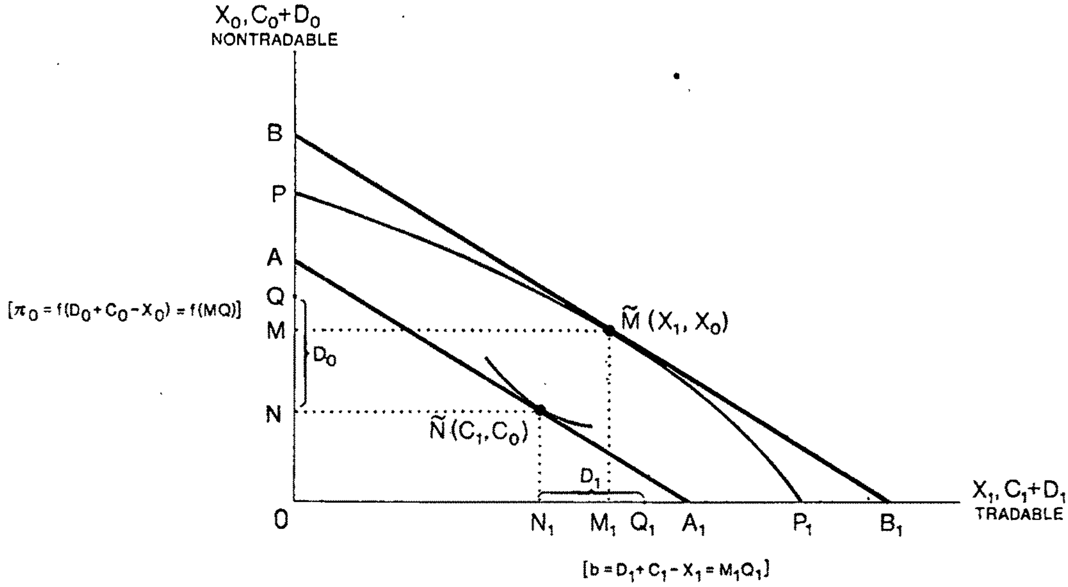



FIGURE 1

and obtain $S_{is}=C_{ic}Y>0$; i.e., there is a positive response to an increase in $s$ for both commodities. Likewise $S_{0q}=X_{0q}-C_{0q}-C_{0c}(1-s)X_1>0$; i.e., the net supply of nontradables unambiguously decreases with an increase in $q$. For tradables, the substitution and income effects work in the opposite directions, and the result is therefore more ambiguous. We get

$$S_{1q} = \frac{X_1}{q}\left(\epsilon_1 + \eta_1 \frac{C_1}{X_1} - \mu_1\right)$$

where $\epsilon_1 = qX_{1q}/X_1$ is the supply elasticity of tradables,

$\eta_1 = -qC_{1q}/C_1$ is the consumption demand elasticity for tradables,

$\mu_1 = (1-s)qC_{1c}$ is the marginal propensity to consume $C_1$ out of $Y$.

Thus $S_{1q}>0$ if and only if

$$\epsilon_1 + \eta_1\left(\frac{C_1}{X_1}\right) - \mu_1 > 0 \tag{7}$$

This formula is, of course, reminiscent of some balance-of-payments stability conditions discussed in the trade literature for the combined classical and Keynesian approaches.[10] However, the present model is somewhat differently structured, emphasizing the tradables/nontradables supply distinction and having no explicit domestic money variable.

To complete the first model in its market equilibrium aspects we now assume the existence of autonomous demands ($D_i$; $i=0, 1$) upon the net supplies ($S_i$) which may in principle consist of public consumption ($G_i$) and private or public investment ($I_i$). For the moment we only look at the short-run equilibrium aspects of inequality between $D_i$ and $S_i$, and leave the possible long-run growth effects of $I_i$ on changes in the $Z$ vector and the *PPC* to Section II.

The fundamental difference between tradable and nontradable goods markets lies in the way in which inflationary (or deflationary) gaps $D_i-S_i$ resolve themselves. As long as the government can control the exchange rate ($e$ or $q$) but foreign

[10] See, for example, Arnold Harberger and Murray Kemp.

borrowing (or investment) is allowed, the market for tradables clears automatically by having net imports equal to the difference $(D_1-S_1)$ if it is positive, or net exports, if it is negative. Denote the outstanding debt of the economy by $B$, its time change by $\dot{B}$, its debt servicing[11] by $R_B$ and the net balance by $b=\dot{B}-R_B$. We can write the adjustment process for tradables, which is also the negative of the balance of payments on current account, in the form:

$$b = \dot{B} - R_B = D_1 + C_1 - X_1 \tag{8}$$

If in Figure 1 the distance $N_1Q_1$ measures $D_1$ we have $b=D_1-S_1$ represented by the distance $M_1Q_1$.

In the market for nontradables no such automatic quantity adjustment can take place. We shall instead assume that any positive gap $D_0-S_0=D_0+C_0-X_0$ (marked by the distance $MQ$ on the $X_0$ axis, Figure 1) will result in pressure on the price level $p_0$, and we can assume a positive relationship between excess demand and the rate of inflation $\pi_0$:

$$\pi_0 = \frac{\dot{p}_0}{p_0} = f(D_0 + C_0 - X_0) \tag{9}$$

where $f'>0$ and $f(0)=0$.[12]

The comparison of the different adjustment processes in the two markets underscores the importance of the *composition* of autonomous demand changes, such as investment or public consumption, in terms of the derived demand in the tradable and nontradable goods markets. Thus suppose for a moment that $D_0=S_0$ so that $p_0$ remains constant, and that there is an increase in $D_1$. This will result in an equivalent increase in the balance-of-payments deficit $(\Delta b=\Delta D_1)$, leaving the real exchange rate $(q)$ and the market for nontradables unchanged. Thus the points $\tilde{N}$ and $\tilde{M}$ will remain stationary. On the other hand, if there is an increase in $D_0$, say, $\Delta D_0$ (from the balance point $D_0=S_0$, i.e., $MQ=0$), $p_0$ will start increasing. This in itself will decrease $q$ in the next period unless the government at the same time corrects the nominal rate $(e)$ so as to keep $q$ stationary (i.e., by making $\dot{e}/e=\pi_0$). A reduction in $q$ would now bring about an accompanying change in the equilibrium output $(\tilde{M})$ and consumption $(\tilde{N})$ levels and the balance-of-payments deficit will increase (assuming $S_{iq}>0$). Likewise $S_0$ will increase, thus making for a subsequent self-correcting reduction of pressure on $p_0$ in the market for nontradables. The system will eventually end up at a higher stationary $p_0$ (and lower $q$, by the amount $\Delta q=\Delta D_0/S_{0q}$). The cumulative indirect change in the deficit brought about by the initial change in $D_0$ will be given by the derivative

$$\frac{\Delta b}{\Delta D_0}\cdot\frac{1}{q} = 1 + \frac{(1-S)X_1 - C_1}{S_{0q}} \lesseqgtr 1$$

iff $(1-S)X_1 \gtrless C_1$, while $\Delta b/\Delta D_1=1$. It follows that under fixed exchange rates, if $(1-S_1)X_1>C_1$, the higher the share of nontradables, in total excess demand, the more of the inflationary pressure will show itself in $p_0$ increases (and $q$ decreases), and the less in an increase in the balance-of-payments deficit (see also Helpman). This type of analysis also explains why under a demand inflation and sticky exchange rates one would expect the domestic price of nontradables (for example, services, construction) to rise by more than that of tradables (such as agricultural products and manufac-

[11] If, as we shall assume below, the economy faces an upward sloping supply curve of foreign loans at marginal interest cost schedule $r(B)$ we have $R_B=\int_0^B r(x)dx$ (thus $dR_B/dB=r(B)$), (see Section II). Note that $R_B$ is part of $GDP$ and $GNP=GDP-R_B$. It does not affect private behavior directly if handled out of the government budget.

[12] We are here confining ourselves only to *ex ante* inflationary (or balance) situations for which $D_0-C_0\geq 0$. *Ex post*, of course, this market must balance. The model could be expanded to cover the case of Keynesian unemployment by assuming $f=0$ for the case of excess supply $(D_0-C_0<0)$, introducing a minimum wage constraint, and allowing for less than full employment of $L$.

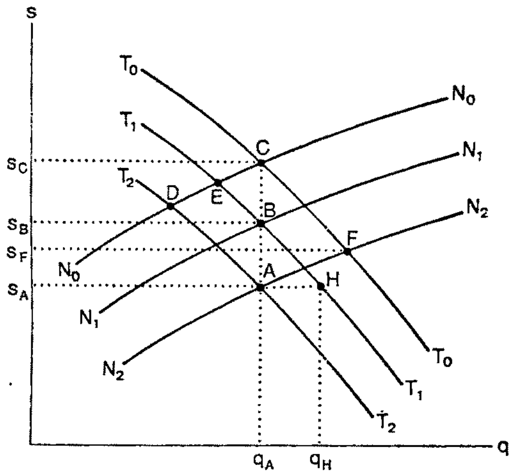



FIGURE 2

tures).[13]

We now use the same set of equations (8) and (9) to elaborate on the interplay of the two basic policy tools, exchange rate policy ($e$ or $q$) and aggregate absorptive measures, summarized under the parameter $s$. For a given level of autonomous demand for tradables ($D_1$), $b$ will decrease if $s$ is increased (since $S_{is}>0$) and if (7) holds it will also be negatively related to an increase in the real exchange rate $q$. The line $T_2T_2$ in Figure 2 represents the various $s-q$ policy combinations keeping $b$ a positive constant. The line $T_1T_1$ represents the same for a lower $b$. The line $T_0T_0$ is drawn for the special case of current account balance (i.e., $\dot{B}=0$). The slope of these lines, which measures the tradeoff between $q$ and $s$, is $[(\partial s/\partial q)_{|b}]=-S_{1q}/S_{1s}$. This is negative if and only if $S_{1q}>0$. If the economy is initially at point $A$ on $T_2T_2$ in Figure 2, a reduction of the deficit to the level corresponding to $T_1T_1$ can be effected, *ceteris paribus*, by pure absorptive measures, keeping $q$ constant at $q_A$ and raising $s$ from $s_A$ (at $A$) to $s_B$ (at $B$) or by exchange rate policy, for given $s$, increasing $q$ from $q_A$ to $q_H$ (at $H$), or by any suitable combination of $s$ and $q$ somewhere between the points $B$ and $H$. However, the nature of the *ceteris paribus* assumption here is critical, because of the simultaneous developments in the other market.[14]

From considerations similar to the above we know that for given autonomous $D_0$ the rate of inflation ($\pi_0$) will fall with an increase in $s$ (since $\partial\pi_0/\partial s=-f'S_{0s}<0$) and rise with an increase in $q$ ($-f'S_{0q}>0$). A devaluation is always inflationary unless coupled with absorptive measures. In analogous fashion the line $N_2N_2$ in Figure 2 represents $s$ and $q$ combinations keeping $\pi_0$ a positive constant. A lower rate of inflation involves a movement of the curve to the left ($N_1N_1$). The curve $N_0N_0$ assumes zero inflation with supply and demand of nontradables balanced at the price level $p_0^*$, say. The slope of the lines $NN$ is always positive since $(\partial s/\partial q)_{|\pi_0}=-S_{0q}/S_{0s}>0$.

The points at which the respective $TT$ and $NN$ lines cross represent pairs of $q$ and $s$ policy combinations that momentarily but simultaneously satisfy the given constraints in terms of allowable inflationary gaps. Moving from $A$ to $B$ and to $C$ here represents a special case of pure absorptive

[13] Let us note that this framework could also be modified to take care of the case in which the market for tradables must balance and we have a flexible rather than a controlled nominal exchange rate. This could be done by replacing equation (8) by an excess demand-price adjustment equation $\dot{e}/e=g(D_1+C_1-X_1)$ where $g'>0$, $g(0)=0$, etc. Alternatively, here as well as for nontradables, domestic money creation could be introduced explicitly whenever there is an *ex ante* gap between $D_i$ and $S_i$. However, the explicit introduction of money (and a real-balance effect) would not change the main aspects of the model for our purpose and they are thus ignored.

[14] Suppose there is a devaluation of $e$ so that $q$ first moves from $q_A$ to $q_H$ (the deficit falling accordingly), and $s$ and $D_i$ are kept constant at their initial values (assuming initially $p_0$ is constant). The price of nontradables ($p_0$) will now start increasing. This process will come to a halt once $p_0$ has increased by the same proportion as $e$. Now $q$ will be back at $q_A$, the deficit will be back at its old level ($T_2T_2$) and all that has happened is an equal rate of increase in the nominal $p_i$ levels, leaving the *relative* price $q$ constant.

policy, keeping $q$ constant (at $q_A$) and affecting both markets simultaneously in the prescribed fashion, by raising $s$ from $s_A$ to $s_B$ and $s_C$. Note that the points $A$ and $B$ (unlike $C$) cannot be sustained unless the nominal exchange rate is being changed at the same rate as the corresponding rate of inflation $\pi_0$ (otherwise $q$ cannot maintain a constant level). Thus strictly speaking a hypothetical movement from $A$ to $B$ really represents a change in the sense of comparative dynamics, switching $\dot{e}/e$ all at once from one steady state to another, at a lower required speed of adjustment.

In the analysis that follows it will be useful to simplify matters by assuming the market for nontradables to remain in balance, say along the line $N_0N_0$, at a constant price level $p_0^*$. In that case a level of $q$ can be maintained by a suitable stationary level of $e$. The combinations of $s$ and $q$ corresponding to successively falling balance-of-payments deficits are represented by the points $D$, $E$, $C$. Note that both $s$ and $q$ must in this case be increased simultaneously because of the need to do away with the accompanying inflationary gap in the market for nontradables. However, a similar effect could alternatively be obtained with a smaller change in $s$ (or none at all) and a larger change in $q$, provided $D_0$ is reduced through either fiscal or monetary measures.[15] The implications of all this for the developments and required policy measures to accompany exchange rate changes (for example, under a "crawling peg") should be clear.

[15] Suppose we start at point $D$ and now move to the point $A$ while reducing $D_0$ to the amount required to keep the nontradable market in balance. This implies that the $N_2N_2$ line is now the line corresponding to constant $p_0$ (and could now be termed $N_0N_0$). For the balance of payments to remain in state $T_2T_2$, $s$ must fall from $s_D$ to $s_A$ while $q$ increases from $q_D$ to $q_A$. Now raise $s$ and $q$ from the point $A$ to the point $F$ at which both markets balance completely. The overall effect is a balance-of-payments correction using a larger increase in $q$ (from $q_D$ to $q_F$ instead of $q_C$) while $s$ falls or increases by less than it would have otherwise ($s_F < s_C$).

## II. Development of the Open Economy

So far we have confined ourselves to price and balance-of-payment changes in the static economy. The incorporation of a growth process into the model can be done by identifying part of the autonomous demands $D_i$ with *investment* in tradables (for example, machines) $I_1$, and/or nontradables (for example, structures) $I_0$, leading to changes in the stocks of assets $K_i$:

$$(10) \qquad \Delta K_i = I_i - h_i K_i \qquad (i = 0, 1)$$

where $h_i$ are the respective depreciation rates.

A change in the stocks of assets will affect the system through the shift vector $Z$, moving the production possibility curve from one "short term" to the next. The precise way in which $Z$ and $PPC$ change depends on a particular specification of the model, for example, whether capital is distributed "optimally" or not. Thus the elements of $Z$ may consist not only of the sum of $K_0$ and $K_1$ but of its particular short-term distribution among industries or firms. For the moment all that we have to assume is that there is some mechanism that makes for $K_i$ formation and allocation and thus for a suitable shift in $PPC$ through the catch-all shift vector $Z$.

We now have to consider $Z$ as a variable together with $q$ in the determination of both outputs $X_i$ and net supplies $S_i$. Denoting by $\Delta$ a total shift from one period to the next and by $\Delta_Z X_i = X_{iZ} \cdot \Delta Z$ (scalar product)[16] a supply shift in $X_i$ due to the long-run factors $Z$ (holding $q$ and $s$ constant) and similarly for $\Delta_Z S_i$ we get:

$$\begin{aligned} \Delta_Z S_i &= S_{iZ} \cdot \Delta Z = X_{iZ} \cdot \Delta Z \\ &\quad - C_{ic}(1-s)(X_{0Z} + qX_{1Z}) \cdot \Delta Z \\ &= \Delta_Z X_i - C_{ic}(1-s)(\Delta_Z X_0 + q\Delta_Z X_i) \end{aligned}$$

[16] $$\Delta_Z X_i = \Delta K_1 \frac{\partial X_i}{\partial K_1} + \Delta K_0 \frac{\partial X_i}{\partial K_0} + \Delta L \frac{\partial X_i}{\partial L} + \Delta t \frac{\partial X_i}{\partial t} \qquad (i = 0, 1)$$

where partial differentiation is performed holding $q$ and $s$ constant.

Denote

$$a = \frac{q\Delta_Z X_1}{\Delta_Z X_0}$$

$$\mu_1 = C_{1e}(1-s)q \qquad \text{(as before)}$$

and $\mu_0 = C_{0e}(1-s)$

where $\mu_i$ are the *MPC* out of gross income and $\mu_0+\mu_1=(1-s)$. We now have

$$\Delta_Z S_1 \gtrless 0 \qquad \text{iff } a \gtrless \frac{\mu_1}{1-\mu_1}$$

$$\Delta_Z S_0 \gtrless 0 \qquad \text{iff } a \lessgtr \frac{1-\mu_0}{\mu_0} = \frac{\mu_1+s}{1-(\mu_1+s)}$$

Since

$$\frac{\mu_1+s}{1-(\mu_1+s)} > \frac{\mu_1}{1-\mu_1} \qquad \text{(as long as } s>0)$$

There are three ranges:

if $\quad a \geq \dfrac{\mu_1+s}{1-(\mu_1+s)},$

$$\Delta_Z S_1 > 0 \quad \text{and} \quad \Delta_Z S_0 \leq 0;$$

if $\quad \dfrac{\mu_1+s}{1-(\mu_1+s)} > a \geq \dfrac{\mu_1}{1-\mu_1},$

$$\Delta_Z S_1 \geq 0 \quad \text{and} \quad \Delta_Z S_0 > 0;$$

if $\quad \dfrac{\mu_1}{1-\mu_1} > a,$

$$\Delta_Z S_1 < 0 \quad \text{and} \quad \Delta_Z S_0 > 0.$$

In particular we note that there will be a secular improvement in the balance of payments for given $s$ and $q$, if and only if:

$$(11) \qquad a > \frac{\mu_1}{1-\mu_1}$$

In words, for given $q$ and $s$, the relative marginal supply shift in tradables over time must be greater than the relative marginal consumer demand for tradables out of income. Note that it is the ratio of the *absolute* size of the increments of $X_1$ and $X_0$ that matters. It will in general depend on past investment and labor training policy, on relative capital intensity in the production of the various goods, and on possible differences in technical progress between the two groups of industries.

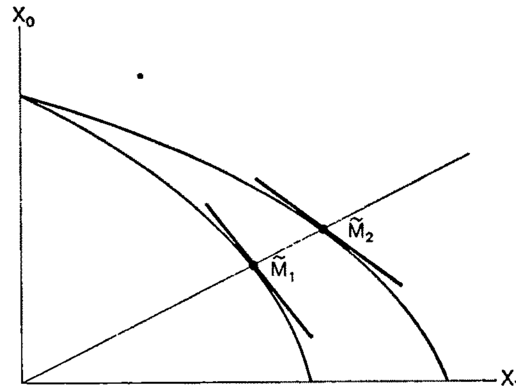


FIGURE 3

A change in $Z$ that is biased toward tradable goods (say, technical progress or investment is biased in that direction or else export prices have increased) will shift the *PPC* so that along a constant relative-output ray from the origin (see Figure 3, movement from $\tilde{M}_1$ to $\tilde{M}_2$) the slope of the *PPC* ($q$) may fall over time, i.e., less of an increase in $q$ will be needed to effect a given *relative* increase in $X_1$.[17] For the analysis of Figure 2 the implication of changes in $Z$ is in terms of a shift of the $TT$ curves to the right and $NN$ curves to the left, with the relative shift depending on the conditions stated above.

We now have the ingredients for a description of development paths in an open economy. At any moment of time $Z$ and *PPC* are given as well as $p_0$, $P_1$, $B$. The choice of $D_i=G_i+I_i$ ($i=0$, 1), $s$, and $q$ (or $e=qp_0/P_1$, assuming government control of the exchange rate) affects the equilibrium in both commodity markets (i.e., $\pi_0$ and $b$) and thus also the domestic

[17] One should be careful in the interpretation of these statements since a rise in $q$ in one period may by itself make the production of tradable goods more profitable and shift factors in that direction in the longer run. Thus the role of $q$ may be underestimated. This is important for empirical estimation of supply responses.

price level $p_0$ and the level of outstanding debt (or exchange reserves) $B$ for the next period. Investment $(I_i)$[18] plus the other factors affecting $Z$ (the amount of exogenously given labor $L$, say, and technical progress) determine a new $PPC$. Now we can start the same process again in the next period.[19] Obviously there is room for many variations on the basic theme, but in principle, this kind of framework could be used to analyze balance-of-payment developments and structural change in a single country over time or as a basis for making comparative analyses of output composition, savings, and balance of payments in different economies.

As an example of explaining international differences in growth patterns consider the characterization of variations in capital inflow across countries. Application of the analysis of the last section suggests that for given $D_1$ and $Z$ we would expect higher $b$ levels to go together with lower $s$ or lower $q$ or both. If extra aid were merely used as a means to finance imports of capital goods $(I_1)$ or defense $(G_1)$, say, variations in $b$ would not be reflected in variations in $S_1$ (since $(D_1-b)$ would not vary). However, one feature of an aid-receiving country should be its ability to smooth its consumption path over time.[20] We would thus expect countries receiving more aid to have, *ceteris paribus* (i.e., for same size and income), a lower savings rate.

One would likewise expect, as long as basic demand patterns in different countries are similar, that countries at the same income level and of the same size, but with a higher relative capital inflow (or the *same* country at a different point in time), would place a lower value on their $q$, i.e., a larger relative share of total production would be in nontradable goods, such as services and construction.

Finally, this framework can be used to rationalize systematic changes in real exchange rates accompanying structural change and balance of payments and capital movements in individual countries over time.

At this stage let me draw attention to another aspect which is of importance when one thinks of this type of framework as a basis for policy. We have so far considered only static optimization. Decisions about savings and investment, foreign borrowing, and the exchange rate were assumed to be taken separately in each period. We know that unless the degrees of freedom on policy making are highly restricted, such behavior will not be optimal from an intertemporal point of view. The ability to dissave and borrow tradable goods first, invest and expand exports (or import substitutes) to repay debts later, must also express itself in proper asset and foreign exchange pricing over time.

In the Appendix I state in greater detail the structure of an optimum intertemporal model. It incorporates the basic assumptions of the general framework under a simple Ramsey-type discounted utility maximand $[\int_0^T e^{-\alpha t}U(C_0, C_1)dt]$ plus a given supply schedule for the foreign debt with marginal cost $r(B)$. The initial and terminal stock levels of capital and debt are given. The solution of such a model as an optimum control problem consists, as one would expect, of two parts. One is a restatement of the static equilibrium conditions under which the marginal rate of substitution in consumption $(U_1/U_0)$ must at each point in time equal that in production $[(\partial F/\partial X_1)/(\partial F/\partial X_0)]$, and both equal the real exchange rate $q$. The second set of relationships is familiar from the theory of optimum capital accumulation (or opti-

[18] It is here assumed that $I_i$ is *ex post* investment, requiring some assumption, at least for nontradables; either $D_0=S_0$ or an *ex ante* imbalance between $D_0$ and $S_0$ is resolved by *ex post* adjustment of $G_0$.

[19] We here identify the "next" period as being the same for investment gestation and for the price change. This, of course, is a simplifying assumption which is not of crucial importance here.

[20] As we shall see below this could also be explained in terms of optimum intertemporal behavior.

mum portfolio choice). If the own net rate of return on each type of asset is denoted by $\rho_i$ ($i=0$) for nontradables and $i=1$ for tradables), then $\rho_i$ measures simultaneously both the marginal net productivity of the capital asset $K_i$ in production as well as the intertemporal rate of substitution of $C_i$ now for a $C_i$ increment in the future. In addition we have:

$$\rho_1 = r(B) \tag{12}$$

$$\frac{\dot{q}}{q} = \rho_0 - \rho_1 = \rho_0 - r(B) \tag{13}$$

These relationships are easy to rationalize intuitively. Equation (12) says that the rate of return on a tradable asset must equal the marginal cost of foreign borrowing. Equation (13) says that at the margin of choice of investing in a tradable asset or in a nontradable asset (or saving for the consumption of nontradables) intertemporal indifference requires the relative capital gain to change at a rate that just equals the difference in the net rates of return of the two assets.

Full analytical characterizations for general models are hard to obtain. On the other hand one can compute numerical solutions for empirically based multisector models having the same basic features as the more general model.[21] However, even in the absence of a full-fledged model the rationale behind these models can be of considerable practical use. For example, the rationale of the more general equation (13) can be used for the choice of a suitable interest rate for public investment decisions.

It is quite common practice to take an estimate of $r(B)$ to represent the accounting rate of interest for public investment decisions. It is a readily available figure and foreign loans naturally represent marginal investment funds. However, as the above theoretical analysis shows, this is definitely a wrong procedure when $q$ is expected to change systematically over time. The correct approach, if one knows the country's balance-of-payments objectives and the relevant supply elasticity of tradable goods, is to get an estimate of expected $\dot{q}/q$. Knowing the marginal cost of foreign borrowing $r(B)$ one can then get an imputed value of the real domestic rate of interest from the sum $\dot{q}/q+r$.[22]

Finally let me point out that the same general framework could also be applied for the rationalization of optimum departures from uniform real exchange rate movements over time. One example for which this can be shown is the case of infant industries or learning effects in the production of some tradables. This calls for a special subsidy to producers. The other is the case of a minimum consumption constraint on some tradables (say low income consumption of a staple food). This calls for a subsidy to the consumer, i.e., an effective exchange rate that lies below the common $q$.

Obviously there are numerous other complications that could be introduced into this basic model. All of this should not distract us from what may be the main implication of the present approach. Far too often exchange rate policy is still seen as a short-term balance-of-payments correcting device and merely as a correction to domestic inflation. This is no doubt strongly tied with the main traditional role of nominal exchange rate adjustments

[21] A detailed programming model that illustrates the application of this kind of framework for Israeli data was given by the author, Dougherty, and Fraenkel. Fraenkel has further developed this model and its theoretical implications, working out time paths for alternative specifications of discount rates and the ability to borrow abroad. His model also provides explicit numerical estimates of the loss of welfare when there exist *absolute* limits on the ability to borrow, a problem no doubt faced in many situations.

[22] In the case of Israel, for example, reasonable orders of magnitude are $r(B)=6$–$8$ percent, $\dot{q}/q=4$–$5$ percent, and thus the real domestic rate of interest (in terms of real domestic resources) should be 10–13 percent.

in well-developed industrialized countries. However, in the process of rapid structural change of an open economy the long-run allocative role of systematic changes in the real exchange rate system may be of much greater importance.

## Appendix

### *The General Optimization Model*

We look at the following general problem:

$$\text{(A1)} \qquad \text{Max} \int_0^T e^{-\alpha t} U(C_0, C_1) dt$$

subject to:

$$\text{(A2)} \qquad F(X_0, X_1; K_0, K_1, L, \text{t}) = 0$$

$$\text{(A3)} \qquad X_0 = C_0 + I_0$$

$$\text{(A4)} \qquad \dot{B} = R_B + C_1 + I_1 - X_1$$

$$\text{(A5)} \qquad \dot{K}_i = I_i - h_i K_i \qquad (i = 0, 1)$$

$$\text{(A6)} \qquad R_B = \int_0^B r(x) dx$$

and $K_i(\text{t})$ $(i=0, 1)$, $B(\text{t})$ are given at $\text{t}=0, T$. The value of $L$ is given at all t.

The formal solution, using the Maximum Principle (see, for example, Robert Dorfman, Kenneth Arrow and Mordecai Kurz), consists of the following necessary conditions:

*Static Maximization:* At *each* point in time the following Hamiltonian must be maximized:

$$\text{(A7)} \quad \text{Max } H = e^{-\alpha t}(U + p_0 \dot{K}_0 + p_1 \dot{K}_1 - p_1' \dot{B})$$

subject to the constraint (A2) where $p_0$, $p_1$, and $p_1'$ are the imputed stock prices, and $H$ can be interpreted as the discounted net domestic product (plus consumers' surplus).[23]

Adding $e^{-\alpha t}\lambda F$ to (A7) with $\lambda$ as a Lagrangian multiplier and substituting from the various equations we must therefore maximize the expression ($H_1$ is the Lagrangian or extended Hamiltonian):

$$e^{\alpha t} H_1 = U(C_0, C_1) + p_0(X_0 - C_0 - h_0 K_0) + p_1(I_1 - h_1 K_1) + p_1'(X_1 - I_1 - C_1 - R_B) + \lambda F(X_0, X_1; K_0, K_1, L, \text{t})$$

with respect to $C_i$, $I_1$, $X_i$ (at given $K_i$, $L$, t).

The first-order conditions are:

$$\text{(A8)} \qquad U_0 = p_0, \qquad U_1 = p_1 = p_1',$$

$$p_0 + \lambda \frac{\partial F}{\partial X_0} = 0, \qquad p_1' + \lambda \frac{\partial F}{\partial X_1} = 0$$

Denoting $q = p_1/p_0$ we get:

$$\text{(A9)} \quad q = \frac{U_1}{U_0} = \frac{p_1'}{p_0} = \frac{p_1}{p_0} = \frac{\partial F/\partial X_1}{\partial F/\partial X_0}$$

We thus get the equality of $q$ and the respective marginal rates of substitution. For (A9) to give a true maximum, concavity of $U(C_0, C_1)$ and convexity of $F(X_0, X_1)$ will be assumed.

*Dynamic Price Equations:* The following dynamic equations must hold for the imputed stock prices:

$$\text{(A10)} \qquad \frac{d(e^{-\alpha t} p_i)}{dt} = -\frac{\partial H_1}{\partial K_i} \qquad (i = 0, 1)$$

$$\text{(A11)} \qquad \frac{d(e^{-\alpha t} p_1')}{dt} = \frac{\partial H_1}{\partial B}$$

Written out, equation (A10) gives

$$\text{(A12)} \quad \frac{\dot{p}_i}{p_i} = \alpha + h_i - \frac{\lambda}{p_i} \frac{\partial F}{\partial K_i} = \alpha - \rho_i$$

where

$$\frac{\lambda}{p_i} \frac{\partial F}{\partial K_i} - h_i = -\frac{\partial F}{\partial K_i} \bigg/ \frac{\partial F}{\partial X_i} - h_i = \rho_i$$

are the own net rates of return.

Similarly, since $p_1' = p_1$ (from (A9)) and $dR_B/dB = r(B)$, we get:

$$\text{(A13)} \qquad \frac{\dot{p}_1}{p_1} = \alpha - r(B)$$

It follows that $r(B) = \rho_1$ (see equation (12)) and finally equation (13) in the text is ob-

[23] Add and subtract $\sum U_i C_i$ in (A7). Consumers' surplus is given by $(U - \sum U_i C_i)$, and net domestic product (in utility units) is given by

$$\left[ \sum_{i=0}^{1} (U_i C_i + p_i \dot{K}_i) - p_1' \dot{B} \right]$$

tained from combining (A12) and (A13).

*Sufficiency:* To have sufficient conditions for an optimum (A1), if it exists, the Lagrangian $H_1$ must be concave in $K_i$, $-B$ (for given $p_i$, t). It follows that $F(\ )$ must be concave in $K_i$ and we must also have $d^2R_B/dB^2 = r'(B) \geq 0$. The above set of equations plus the boundary values on the stocks in principle lead to a complete characterization of an optimum program including the pattern of foreign debt, savings, and prices over time.

Finally we note that in all of the above we have ignored the possibility of corner solutions, otherwise the nonnegativity constraints (for example, on $C_i$ and $I_i$) must be taken into account explicitly in the maximization and the corresponding price variables. The fact that we get $p_1' = p_1$, for example, is connected with an implicit assumption that it will not be optimal to reduce $K_1$ (otherwise $p_1'$ and $p_1$ may differ by a suitable Lagrangian multiple).[24] Also, if $p_1 = p_1'$ an optimum solution may very well imply an initial "jump"[25] (i.e., instantaneous adjustment) in one of the two state variables $K_1$ or $B$ so as to get the two rates of return $r(B)$ and $\rho_1$ equalized. This also implies that in the initial and terminal boundary conditions it is the *difference* $B(t) - K(t)$ that mainly matters for characterization of the path rather than the individual values of $B$ and $K$ separately, at $t=0, T$.

[24] Alternatively, we could *assume* that $K_1$ and $-B$ are freely convertible into each other, in which case $\dot{K}_1 < 0$ could occur.

[25] For a mathematical treatment of such a case, see Karl Vind, Arrow and Kurz.

## REFERENCES


**K. J. Arrow and M. Kurz,** *Public Investment, the Rate of Return and Optimal Fiscal Policy,* Baltimore 1970.

**E. Berglas and A. Razin,** "Real Exchange Rate and Devaluation," *J. Int. Econ.*, May 1973, *3*, 179–91.

**M. Bruno,** "Optimal Patterns of Trade and Development," *Rev. Econ. Statist.*, Nov. 1967, *49*, 545–54.

———, "Trade, Growth and Capital," M.I.T. Econ. Dept. working pap. no. 65, 1970.

———, **C. Dougherty, and M. Fraenkel,** "Dynamic Input-Output, Trade and Development," in A. P. Carter and A. Brody, eds., *Applications of Input-Output Analysis,* Amsterdam 1969.

**R. Dorfman,** "An Economic Interpretation of Optimal Control Theory," *Amer. Econ. Rev.*, Dec. 1969, *59*, 817–31.

**R. Dornbusch,** "Devaluation, Money, and Nontraded Goods," *Amer. Econ. Rev.*, Dec. 1973, *63*, 871–80.

**M. Fraenkel,** "Long Term Planning of the Balance of Payments," presented at the Input-Output Conference, Geneva, Jan. 1971.

**A. C. Harberger,** "Currency Depreciation, Income, and the Balance of Trade," *J. Polit. Econ.*, Feb. 1950, *58*, 47–60.

**E. Helpman,** "Macroeconomic Policy and Employment in an Open Economy with Nontraded Goods," unpublished doctoral dissertation, Harvard Univ. 1974.

**M. C. Kemp,** *The Pure Theory of International Trade,* Englewood Cliffs 1964.

**A. O. Krueger,** "Balance-of-Payments Theory," *J. Econ. Lit.*, Mar. 1969, *7*, 1–26.

**P. M. Oppenheimer,** "Non-Traded Goods and the Balance of Payments: A Historical Note," *J. Econ. Lit.*, Sept. 1974, *12*, 882–87.

**I. F. Pearce,** "The Problem of the Balance of Payments," *Int. Econ. Rev.*, Jan. 1961, *2*, 1–28.

**K. Vind,** "Control Systems with Jumps in the State Variables," *Econometrica,* Apr. 1967, *35*, 273–77.

# A Bargaining Theoretic Approach to Cropsharing Contracts

*By* Clive Bell and Pinhas Zusman*

Sharecropping is a geographically widespread and historically tenacious form of lease. The questions of how the rental share is determined, and how such a form of lease affects resource allocation and the distribution of income pose a challenge to economic theory which has evoked two notable recent attempts to provide an explanation. First, within what may be called the Marshallian tradition, P. K. Bardhan and T. N. Srinivasan (1971) (hereafter B-S) have developed a fairly complete scheme of analysis in which the rental share is regarded by the parties to the contract "as a price-like variable" (1974, p. 1068), i.e., it is parametrically given to all agents. For an exogenously given real wage rate, the supply of and demand for leases determine the rental share. A second approach is that originally formulated by S. N. S. Cheung and subsequently extended and generalized by D. M. G. Newbery (1973a, b), (hereafter C-N). In their system, also claimed to be competitive, both the rental share and the minimum labor input per unit of land are stipulated in all contracts, their values being jointly determined by the landlord's desire to maximize his income subject to the condition that the tenant's income does not fall below his alternative earnings in a perfectly competitive labor market.

True to the spirit of Marshall's analysis (p. 644), B-S show that resource allocation under sharecropping departs from that ruling in a competitive system with fixed rent leases, it being well known that the latter is Pareto optimal. As Newbery (1974) has shown, the fundamental difficulty with B-S's formulation is that there will be excess demand for land for all rental shares less than unity, unless the marginal productivity of land is zero. By contrast, in the C-N scheme of things, the equilibrium which obtains happens to be identical with that ruling in a competitive fixed-rent system and is therefore Pareto efficient. This formulation, however, has two basic and related difficulties of its own. First, there is the dubious enforceability of the minimum labor-intensity provision of the contract. Secondly, if land has to be rationed, both agents are placed in a bargaining situation. Hence, one would expect the rental share which emerges from the bargaining process to reflect the agents' respective bargaining strengths.

Once the overriding importance of the last point is recognized, it is plain that the problem must be reformulated and analyzed within the framework of bargaining theory. This is the salient feature of our approach. Now as Newbery and Joseph Stiglitz have shown, the choice of rental contract is intimately related to the allocation of risk bearing and the provision of incentives in an uncertain world. However, in order to keep the bargaining aspects of the problem in sharp focus, we confine ourselves here to pure cropsharing under certainty.

In addition to these theoretical con-

* Economists, Development Research Center, World Bank. Bell is on leave from the Institute of Development Studies, Sussex, and Zusman from the Hebrew University, Jerusalem. We would like to acknowledge helpful comments by Bela Balassa and an anonymous referee. Errors of analysis are ours alone.

siderations, the foregoing competitive formulations pay little attention to the fact that in addition to his labor, the tenant often possesses land of his own, some capital, and entrepreneurial and management skills. In this connection, it is widely observed that landlord and tenant split the produce in proportions which are close to, if not exactly, 50:50. Any analysis which seeks to explain the latter must take account of the former. All this suggests that the assumptions on which both theories rest must be reconsidered, and that the analysis of the problem must be tailored to the characteristics of the agrarian system in question.

In this paper, we discuss first some empirical evidence concerning the existence and nature of nontradable factors and market imperfections, which play a key role in the bargaining process. The evidence relates mainly to northeast Bihar, but it should survive extension to much of the lower Indo-Gangetic plain. Secondly, we employ John Nash's solution to the bargaining problem, and hence derive the equilibrium rental share. By way of illustration, we calculate the predicted rental and factor shares corresponding to some sets of plausible production parameters. Finally, we analyze briefly the situation in which the marginal productivity of land is zero.

## I. Assumptions and Evidence

The most natural question to begin with concerns the specification of the labor market. The assumption of a perfectly certain and elastic supply of labor and work opportunities at an exogenously determined and parametrically given wage is not really acceptable at the microeconomic (village) level. But it will be retained on account of its simplicity, which eases the task of analyzing the effects of relaxing other less defensible assumptions.

First, the relationship between agricultural worker and tenant status is probably asymmetric. It is always open to the tenant to work for others as well as (or instead of) on the land he has leased in; but the landless laborer cannot just elect to become a tenant if that status requires capital and husbandry skills to which he can get access only with great difficulty, if at all. In this connection, it should be noted that in both India as a whole and Bihar, household operational holdings which are partly owned and partly leased greatly outnumber those which are wholly leased in, and that the former account also for the lion's share of all land leased in.[1] Thus, tenants are drawn mainly not from the mass of landless laborers, but from the ranks of the small peasantry possessing land of their own as well as skills and capital (or access to capital), all of which are traded (if, indeed, they are tradable at all) in imperfect markets.

Moreover, the micro data analyzed by Krishna Bharadwaj for several Indian States and by Bell for northeast Bihar reveal that both categories of tenant households make very extensive use of hired labor for cultural labor operations and harvesting. Thus the costs of supervising cultivation itself (though not the enforcement of a minimum input bundle, of course) fall on the tenant. Moreover, as the work is done by casual laborers, who must be contacted in sufficient numbers at the right time, there are transactions costs of labor hire which are passed on to the tenant by the landlord leasing out his land. Factors such as choice of crop, timeliness of operations, and good husbandry, where the scope for them is large, indicate that purely entrepreneurial skills are also involved (see C. H. H. Rao). All these point to a vital managerial element in the tenant's role—for which he must be sufficiently rewarded.

[1] See, for example, the Indian *National Sample Survey* (nos. 36, 113, 144) and S. K. Chakraverty, p. 99.

The second major query surrounding the stylized competitive models is the absence of nonlabor inputs other than land. Here the principal difficulty stems from the nature of the market for animal draught power, especially on the supply side. In India, there is no historical evidence of human traction (or digging) for land preparation (except for small-scale vegetable cultivation), and there is a broad social disdain for such methods (see Prufulla Sanghvi). Now, if the markets for both this input and credit were perfect, Cheung's contract could be extended to stipulate the minimum draught input per acre as well as the rental share and minimum labor intensity. However, the field data suggest that: (a) hired draught is only a very small fraction of total draught inputs, in marked contrast to labor; and (b) owning at least one pair of bullocks or buffalo is a sine qua non for obtaining a lease, and that this distinguishes tenants owning no land of their own from the general class of landless laborers.

These findings are backed up by the comments of tenants during interviews: although there existed a daily rate for the hire of a bullock team season by season, it was extremely difficult in practice to hire. No cultivator could afford to rely on the hire market, nor, indirectly, could any landlord. Hence the likelihood that tenants would own some land, for land and draught livestock ownership are very highly correlated. Nor are such observations confined to Bihar. In his study of allocative efficiency in an Eastern Uttar Pradesh village, David Hopper was unable to compare the calculated marginal product of bullock labor with its market price because "there was not an active local market for bullock rentals to provide a comparison" (p. 623).

The upshot of this examination of the assumptions underlying the competitive models is that the conclusions drawn from them may be highly misleading. In itself, the generalization of those analyses to many kinds of labor and nonlabor inputs presents no difficulties, provided all markets are perfect. But once it is perceived that tenants are distinguished from ordinary landless laborers by their command over factors whose services are traded (if directly tradable at all) in markets which are inherently and highly imperfect, then major revisions become necessary. First, the equilibrium condition that under perfect certainty the tenant's income must be equal to his alternative earnings as a wage laborer has to be dropped. Secondly, it cannot be replaced by a generalized equivalent to include the alternative earnings of factors other than "worker labor" owned by tenants because the markets for such factors are not even remotely perfect. Thirdly, while it may be safe to assume that the market for unskilled labor is perfect, any formulation of the problem must respect the fact that the crux of the land lease decision lies in the nontradability of managerial skill and draught power.[2]

These imperfections are intimately connected with the issue of whether or not the landlord has the ability to enforce contracts specifying a set of minimum input intensities on sharecropped land. In addition to the a priori arguments advanced by B-S that such contracts are not enforceable, there is empirical evidence to that effect for northeast Bihar. The detailed testing of the competing hypotheses advanced by the two schools[3] has been undertaken elsewhere (Bell), with the Marshallians coming off the better. Although there is no space here to go into much detail, the fact that most tenants also own land allows a search for the influ-

[2] In this connection, it is worth noting Richard H. Day's conclusion that mechanization contributed in a major way to the decline of sharecropping in the *U.S.* South by providing an alternative to tenant-owned draught power.

[3] Four central hypotheses are set up and tested by Cheung (pp. 55–61), who uses evidence from prerevolutionary China, Taiwan, Korea, and Japan.

TABLE 1—MEAN INPUTS AND OUTPUTS PER ACRE ON TENANT HOLDINGS

| | Cropping Intensity | Intermediate Inputs[a] | Hired Labor[a] | Yield[a] |
|---|---|---|---|---|
| Sharecropped plots | 1.40 | 33.0 | 35.6 | 378.3 |
| Owned plots | 1.76 | 64.3 | 46.7 | 561.9 |

*Source:* Bell.
[a] Rupees per net acre sown.

ence exerted by nontradables and other imperfections on the allocation of resources. For each such tenant, his access to inputs (traded or nontraded) and his aversion to risk are common elements in his decisions concerning the allocation of resources between owned and sharecropped plots. These sources of variation across farmers can be eliminated in the following way: for each farmer, form a vector whose elements are the differences between output and inputs per acre on owned and sharecropped land, respectively. This procedure amounts to a "pairing" of observations, a classic and powerful experimental method which has been neglected in the sharecropping literature. Table 1 summarizes mean input and output levels by the tenure status of land cultivated for a sample of 31 farmers, cropping intensity being a proxy for land services (and hence, indirectly, for draught services). Given the multivariate nature of the situation, an appropriate way of testing the significance of the vector of differences on owned and sharecropped land, namely [0.36, 31.3, 11.1, 183.6], is to employ Hotelling's $T^2$ statistic. The computed value of $T^2$ is 31.37, which is significant at the 0.1 percent level.

These findings suggest that the tenants were highly successful in "diverting" both nontradable and tradable resources to the land which they owned, contrary to the competitive theory offered by C-N. While landlords may have had a mind to enforce contracts specifying the mean vector of input intensities and yields for owned land, the task was apparently beyond them in practice.

## II. Bargaining Over the Rental Share

The foregoing discussion suggests a number of fruitful departures from competitive assumptions. We shall confine our attention to just two of them: the inclusion of nontradable factors and the rationing of leases, by landlords. These necessitate the use of a quite different analytical approach, namely, a bargaining game formulation. To keep matters simple, we assume that tenants are landless, but endowed with managerial and husbandry skills, and draught power. The introduction of the tenant's own land would add the complications of an allocation problem to those of finding a solution to the bargaining game.

Suppose just one landlord and $n$ $(<\infty)$ identical tenants reside in each locality and all land is homogeneous. The landlord, who does not cultivate any land himself,[4] possesses $\alpha$ units of land, while each tenant household is endowed with one unit of labor. The labor market is perfect so that the marginal utility of leisure equals the wage rate ($w$) and tenants maximize income net of the imputed cost of labor. Contracts specifying minimum intensities are unenforceable so the bargaining process will determine the rental share ($r$) as the sole provision of the contract. With the existence of nontradable factors, it is reasonable to assume that each tenant's output is a strictly concave, twice continuously differentiable function of his land holding($h$) and unskilled labor inputs ($l$).[5]

[4] This assumption is made for simplicity's sake. The value of the landlord's leisure is then independent of the number and nature of the contracts he concludes, and therefore will not feature in the analysis.

[5] For simplicity, fully traded intermediate inputs such as seed, fertilizer, etc. are lumped together with unskilled labor. Table 1 indicates that this procedure is also empirically defensible.

Consider the landlord's dealings with a single tenant in the situation in which contracts with all other tenants have been closed at a rental rate $\bar{r}$. If these two agents agree on a contract specifying a rental share $r$, their respective incomes (including the value of the tenant's leisure) will be:

$$(1) \quad \begin{aligned} Y^1 &= (1-r)F(h, l_1) - wl_1 + w \\ Y^2 &= (n-1)\bar{r}F[(\alpha-q)/(n-1), l] \\ &\quad + rF(h, l_1) \end{aligned}$$

where $l_1$ is the total labor input (i.e., family plus hired hands) of the tenant in question, and $h \leq q$, $(\alpha - q)$ being the area of land leased out to the other tenants and, because they are identical, equally divided among them. Now suppose that the marginal product of land is positive ($F_1 > 0$) so that there is excess demand for tenancies ($r < 1$), and land is fully utilized ($q = h$) and rationed. Maximization by the single tenant operating $h$ units of land, and by all other tenants, implies also that

$$(2) \quad (1-r)F_2(h, l_1) = (1-\bar{r})F_2[(\alpha-h)/(n-1), l] = w$$

respectively. If, however, the two agents fail to agree, the tenant will fall back on wage labor and the landlord will distribute the $h$ units of land thus freed among the other $(n-1)$ tenants, who change their labor inputs accordingly. In this situation, the tenant's and landlord's incomes become, respectively:

$$(3) \quad \begin{aligned} y^1 &= w \\ y^2 &= (n-1)\bar{r}F(\alpha/(n-1), l') \end{aligned}$$

where the labor input on each of the $(n-1)$ tenancies is now given by $(1-\bar{r})F_2(\alpha/(n-1), l') = w$. The disagreement payoffs ($y^1$ and $y^2$) thus represent the bargaining power of the tenant and the landlord, respectively. The solution of the bargaining problem must reflect the respective bargaining positions of both parties. In this connection, it should be noted that in the event of a disagreement, the landlord's ability to distribute any tenant's parcels among the other claimants makes his position close to what it would be if he were to engage in cultivation himself. Indeed, at the margin, he would be indifferent between such a redistribution and a little extra self-cultivation.

We now employ John Nash's method to arrive at a solution to this bargaining game.[6] Define

$$(4) \quad N = (Y^1 - y^1)(Y^2 - y^2)$$

Then any value of $r$ which maximizes $N$ is a solution. However, since labor utilization by tenants is not part of the contract, tenants' behavior, as given by equation (2), constitutes an external constraint on the bargaining game. That is, the influence that the rental share exerts on labor utilization is recognized by both parties and affects the outcome of the game. Nash's solution also requires that $Y^i - y^i > 0$ for $i = 1, 2$. That is, neither the tenant nor the landlord will enter into nonremunerative contracts. Substituting (1) and (3) into (4), we get the following first-order condition for $N$ to be a maximum:

$$(5) \quad \begin{aligned} \frac{\partial N}{\partial r} = &- F(h, l_1)\{rF(h, l_1) + \bar{r}(n-1) \\ &\cdot [F(h, l) - F(\alpha/(n-1), l')]\} \\ &+ \left[F(h, l_1) + rF_2\frac{dl_1}{dr}\right] \\ &\cdot [(1-r)F(h, l_1) - wl_1] = 0 \end{aligned}$$

where, from (2)

$$\left.\frac{dl_1}{dr}\right|_{w,h} = \frac{F_2}{(1-r)F_{22}} \leq 0$$

[6] In Nash's solution of a two-player cooperative bargaining game, the agreement payoffs to the parties depend on their "disagreement payoffs" and the feasible set of all payoffs. The solution is consistent with the following four axioms: (i) "group rationality"; (ii) von Neumann invariance of the utility indicators; (iii) "independence of irrelevant alternatives"; and (iv) symmetry of players. Alternatively, Nash's solution may be viewed as the outcome of some plausible bargaining processes (see R.D. Luce and H. Raiffa).

Thus, any $r^0$ satisfying (5) is a function of $\bar{r}$. That is, the rental share negotiated in a particular contract depends on the share stipulated in the other $(n-1)$ contracts. The equilibrium configuration of contracts is such that if any contract expires, all other contracts being given, renegotiation will yield the very same contract. Recalling that by assumption all tenants are identical, then symmetry considerations imply $r^0=\bar{r}$. Hence, by (2), $l_1=l$ and $h=\alpha/n$ ( implying, as one would expect, that all tenants would cultivate parcels of equal area). Thus, from (5), we have

$$(6)\quad r^0=[1-\beta_2]\Bigg/\left\{(n+1)-(n-1)\frac{F(')}{F(\cdot)}-\beta_2[1+\epsilon_2(1-\beta_2)]\right\}$$

where $\beta_2$ is the point elasticity of output with respect to labor; $F(\cdot)=F(\alpha/n,\, l)$ is the output on each farm in the event of agreement; $F(')=F(\alpha/(n-1),\, l')$ is the output on each of the other $(n-1)$ tenancies if there is a disagreement; and $\epsilon_2=F_2/l_1F_{22}$ (the inverse of the elasticity of the marginal productivity of labor). Thus equation (6) provides a general determination of $r^0$ once $\alpha$, $n$, $w$, and the technology are specified.

In order to illuminate the influence exerted by these factors on the rental share, we shall now derive solutions for technologies in which the elasticities of substitution between land and unskilled labor are zero, unity, and infinity, respectively.[7]

(A) *Zero elasticity of substitution technology*: $F=\min\,[h^\nu, l^\nu]$, $0<\nu\leq 1$. As tenants are maximizers, $h=l$. Without loss of generality, let $\alpha=n$ so that only family labor is used. By writing down the agreement and disagreement payoffs and maximizing $N$, it may be shown that in a universal cooperative solution, $n=\alpha$, $h^0=l_1^0=1$, and

$$(7)\quad r^0=(1-w)/\{2+(n-1)[1-(1-1/n)^{-\nu}]\}$$

which, for $n$ sufficiently large, becomes

$$(7')\qquad r^0=(1-w)/(2-\nu)\,.$$

Now for the lease contracts to be remunerative, $w<1$ (the level of output on each tenancy). Thus $(1-w)$ is the "surplus" to be divided between the tenant's nontradables and the landlord's land. If there are numerous tenants and constant returns to scale in land and unskilled labor alone $(\nu=1)$, $r^0=1-w$, which is the C-N solution. The competitive share of nontradables is $(1-\nu)$, so that the competitive rental share is $(\nu-w)$, which is less than $r^0$ from (7′) for $\nu<1$, the difference reflecting the landlord's bargaining power.

An important implication of a fixed coefficient technology is that there is no incentive element in the contract, as labor inputs per acre are fixed. The landlord has no enforcement problem, and $r$ acts purely as a distributive instrument. Under these production conditions, it follows immediately that

$$dl_1/dr\,\big|_{w,h}=0$$

Hence, from (3) and (5), $Y^1-y^1=Y^2-y^2$, i.e., the agents' gains over their respective disagreement payoffs are equal.[8]

(B) *Cobb-Douglas technology*: $F=Ah^{\beta_1}l^{\beta_2}$, $A$ constant, $0<\beta_1+\beta_2\leq 1$. Here, consider the limiting case when $\alpha/n(n-1)$ —the extra area received by each of the other $(n-1)$ tenants if the $n$th fails to reach a contract—is small.[9] Then, by expanding $F(')$ as a Taylor series about $F(\cdot)$, and using the condition that the other

[7] It turns out that the derivation using a general *CES* technology does not yield simple interpretable expressions.

[8] For this production technology, the utility possibility frontier $U(Y^1, Y^2)$ for given $\alpha$, $n$, and $w$ is linear.

[9] For example, in a village possessing 1000 acres and 100 tenant households, $\alpha/(n-1)n\simeq 0.1$ acres.

TABLE 2—VALUES OF $r^0$ FOR NUMEROUS TENANTS AND COBB-DOUGLAS TECHNOLOGY

| $\beta_1$ | 0 | 0.2 | 0.4 | 0.6 |
|---|---|---|---|---|
| $\beta_2$ | | | | |
| 0 | 0.50 | 0.56 | 0.62 | 0.71 |
| 0.2 | 0.40 | 0.46 | 0.53 | 0.64 |
| 0.4 | 0.30 | 0.36 | 0.45 | 0.60 |
| 0.6 | 0.20 | 0.27 | 0.40 | — |

tenants alter their labor inputs on receiving an extra $\alpha/n(n-1)$ units of land so as to maintain $F_2=w/(1-r^0)$, some manipulation gives

$$r^0 = (1 - \beta_2)/[2 - \beta_1/(1 - \beta_2)] \tag{8}$$

which is independent of factor endowments and the wage rate. As $l$ decreases with $r$ for given $w$ and $h$, the rental share must now play an allocative as well as a distributive role.

It is instructive to compute $r^0$ from (8) for various pairs of $\beta_1$ and $\beta_2$. If there are constant returns to scale in $h$ and $l$ ($\beta_1+\beta_2=\nu=1$), (8) reduces to $r^0=\beta_1$ which is the C-N solution. However, resource allocation is not Pareto optimal, for $F_2=w/(1-r^0)>w$ when contracts stipulating a minimum labor intensity cannot be enforced and the elasticity of substitution between $h$ and $l$ is nonzero. Thus, while land receives its competitive cost share, labor inputs per unit of land are lower than they would be under fixed rents, and production is not efficient.

More generally, from Table 2, we see first that $r^0$ exceeds the output elasticity of land by a good margin if $F$ is highly concave. This is noteworthy because although the landlord is a potential monopolist, all land is under cultivation. For the assumptions on the technology ensure excess demand for land by tenants ($\beta_1$ and $\beta_2$ are constant, and $F_1>0$ everywhere). Also, when $F$ is strictly concave, a maximizing landlord will not keep land out of cultivation, since $r^0$ is insensitive to changes in $\alpha$ when $\alpha/n(n-1)$ is sufficiently small, and total output will fall if there is a failure to reach agreement with one or more tenants. Secondly, there is a marked tendency for $r^0$ to take values nearer to the magic 0.5 than $\beta_1$ for virtually all but extreme $(\beta_1, \beta_2)$ pairs. Thirdly, for plausible values of $\beta_1$ and $\beta_2$, $r^0$ clusters closely around 0.5. Hopper, for example, reports values of $\beta_1$ in the range 0.2–0.4 and of $\beta_2$ in the range 0–0.2 (p. 615).

A related issue is the structure of factor shares corresponding to the ruling value of $r^0$. The share of (field) labor, whether supplied by the tenant's own household or by the landless is given immediately by equation (2):

$$s_w = wl/F(\cdot) = (1 - r^0)lF_2(\cdot)/F(\cdot) = (1 - r^0)\beta_2$$

Thus, following the Marshallian rule for the use of the variable factor not only depresses employment, but also reduces labor's share in output. The residual output, of course, accrues to the nontradable factor. Denoting the factor shares of land, field labor, and nontradables by $s_l$, $s_w$, and $s_t$, respectively, the most plausible values of $\beta_1$ and $\beta_2$ yield the sets of factor shares shown in Table 3. Taking the competitive structures of factor shares as a benchmark, it is seen that: (a) the landlord exploits the tenant; (b) both the tenant (as manager) and the landlord exploit labor, though only in the fifth case is the landlord's exploitation of the tenant more than offset by the tenant's exploitation of labor.

(C) *Land and labor perfect substitutes:* $F=(h+l)^\nu$, where $\nu\leq 1$. Here, $F_1>0$ everywhere, so there is excess demand for land if $r<1$. Also, from (2), $\nu(1-r)(h+l_1)^{\nu-1}=w$. Thus, if land is redistributed in the event of a disagreement, the labor input employed by each of the other $(n-1)$ tenants will fall so as to keep $(h+l)^{\nu-1}$ constant. Hence production on the other tenancies will not change following a redistribution

TABLE 3—FACTOR SHARES

| | | | | | | |
|---|---|---|---|---|---|---|
| $\beta_1$ | 0.4 | 0.2 | 0.4 | 0.2 | 0.4 | 0.2 |
| $\beta_2$ | 0 | 0 | 0.2 | 0.2 | 0.4 | 0.4 |
| $s_l$ | 0.62 | 0.56 | 0.53 | 0.45 | 0.45 | 0.36 |
| $s_w$ | 0 | 0 | 0.10 | 0.11 | 0.22 | 0.26 |
| $s_t$ | 0.38 | 0.44 | 0.37 | 0.44 | 0.33 | 0.38 |

of land, and the landlord's bargaining position is the weaker thereby. In this case, the Nash function takes the simple form

$$(4') \qquad N = rF(\cdot)[(1 - r)F(\cdot) - wl_1]$$

from which we obtain, at length,

$$(9) \qquad r^0 = \frac{(1 - \beta_2)(1 - \nu)}{(2 - \nu - \beta_2)}$$

where $0 \leq \beta_2 \leq \nu$ and is also a function of $w$.

To round off the analysis, it is instructive to compare the rental shares ruling in the above cases as $w$ varies, but with given factor endowments ($\alpha$, $n$) and returns to scale ($\nu$). In the zero elasticity of substitution case (7′), it is immediate that $r^0$ is linearly decreasing in $w$, with $r^0 = 1/(2-\nu)$ ($\leq 1$ for $\nu \leq 1$) at $w=0$, and $r^0=0$ at $w=1$ ($w \leq 1$ if lease contracts are remunerative to tenants). Now in a Cobb-Douglas technology, $r^0$ is constant and less than unity for $\nu < 1$. It is easily shown by comparing (7′) and (8) that there exists some $0 < w^* < 1$ such that the rental share ruling in a Cobb-Douglas technology is less than that in a Leontief technology if $w < w^*$, and conversely if $w > w^*$.

The perfect substitutes case is more complicated. Here, note first that by (2), $\beta_2(w) = \nu l/(1+l)$, where $h^0 = 1$ ($\alpha = n$). By definition, $0 < \beta_2(w) \leq \nu$. Also $\beta_2(w) \rightarrow 0$ as $l \rightarrow 0$ and $\beta_2(w) \rightarrow \nu$, its upper bound, as $l$ becomes large. As $\beta_2(w) \rightarrow 0$, $r^0 \rightarrow (1-\nu)/(2-\nu)$, and as $\beta_2(w) \rightarrow \nu$, $r^0 \rightarrow (1-\nu)/2$. By differentiating $r^0$ in (9) with respect to $\beta_2$, we see that $dr^0/d\beta_2 < 0$ for $\nu < 1$ so that $r^0$ decreases monotonically as $\beta_2$ increases from 0 to $\nu$. Now as $r^0 \rightarrow (1-\nu)/2$ for large $l$, it follows from the tenants' maximizing behavior $[\nu(1-r)(1+l)^{\nu-1} = w]$ that this is the limiting value of $r^0$ as $w \rightarrow 0$. Moreover, because $r^0 = (1-\nu)/2$ is the maximum of all $r^0$ on $\beta_2 \epsilon [0, \nu]$, and as the Cobb-Douglas value in (8) exceeds this value, the rental share in technology $C$ is less than that in $B$ for all $w > 0$. A further deduction from tenants' decisions regarding labor utilization is that because $r^0 = (1-\nu)/(2-\nu)$ for $\beta_2(w) = 0$ (which implies $l = 0$), then no labor will be used if $w \geq \nu[1 - (1-\nu)/(2-\nu)] = \nu/(2-\nu) < 1$. In this range of $w$, $r^0 = (1-\nu)/(2-\nu)$. Also, we note from (9) that if $\nu = 1$, then $r^0 = 0$ for all $w \geq 0$, a result also obtained by Newbery (1974, p. 1063). The intuitive interpretation of the behavior of the rental shares in different technologies is that for a given wage rate, the higher the elasticity of substitution, the greater the incentives (lower $r$) to use labor the landlord will provide to the tenant. An increase in the wage rate improves the bargaining strength of the tenant and thus tends to lower the rental share. However, the associated fall (if any) in the latter depends in a complicated way on the elasticity of substitution and the wage rate.

## III. Some Land Uncultivated: Causes and Consequences

To complete the analysis, we now consider the cases which may lead to some land remaining unutilized. At first glance, this may come about in two ways: either all tenants are able to choose a parcel size such that $F_1 = 0$ (assuming an appropriate production function); or the parcel size is

also determined in the bargaining process. Let us deal with each of the cases in turn.

Treading a now familiar path, we have in the former case, if a contract is reached, conditions described by equation (1) above. If there is a failure to reach agreement, then we have

$$(10) \quad y^1 = w, \quad y^2 = (n-1)\bar{r}F(\bar{h}, l)$$

In this case,

$$(11) \quad N = [(1-r)F(\cdot) - wl_1][rF(\cdot)]$$

where $F(\cdot) = F(h, l_1)$. Note that $N$ is now independent of the contracts reached with the other tenants. Recall that labor inputs applied by the tenant depend on the rental share—equation (2)—and that tenant maximization implies $F_1(h, l_1) = 0$. Differentiating $N$ partially with respect to $r$, we get

$$(12) \quad \frac{\partial N}{\partial r} = -rF(\cdot)^2 + \left[F(\cdot) + rF_2 \cdot \frac{dl_1}{dr}\right] \cdot [(1-r)F(\cdot) - wl_1]$$

By differentiating totally the tenant's maximizing conditions, viz., $F_1 = 0$ and (2), and hence solving for $dl_1/dr$, we obtain

$$\frac{dl_1}{dr} = \frac{1}{(1-r)} \cdot \frac{F_2 F_{11}}{\Delta}$$

where $\Delta = F_{11}F_{22} - F_{12}^2 > 0$ by the strict concavity of $F$. Hence, for $N$ to be a maximum simultaneously for all contracts, we have, from (12),

$$(13) \quad r^0 = \frac{(1-\beta_2)}{2 - \beta_2[1 + (1-\beta_2)F_2F_{11}/\Delta]}$$

Now, for some $r^0 > 0$, but arbitrarily small, $[(1-r^0)F(\cdot) - wl_1] > 0$ if land is at all cultivated. Also from (12),

$$\left.\frac{\partial N}{\partial r}\right|_{r=0} = F(\cdot)[(1-r)F(\cdot) - wl_1] > 0$$

for the same reason. Then, since $0 \leq r^0 \leq 1$, the maximum value of $N$, and thus the solution to the bargaining game, exists and is associated with a positive rental share. This despite the fact that the marginal productivity of land is zero.

It is important to relate the analysis of this case to the discussion of the competitive formulations in the literature. Recall that, as Newbery has shown, $F_1 = 0$ is the only condition under which a B-S equilibrium will exist. Given fixed factors in limited supply (in this case, nontradables) together with an abundance of land, it is highly plausible that the marginal product of land will fall to zero if the parcel size is made sufficiently large. This may well be the case in some parts of Latin America. What is striking is that our analysis predicts a positive rental share in this situation. That the rental share is positive stems directly from the landlord's power to prevent a would-be tenant from cultivating land if the landlord so chooses.

The second case is one in which both the size of the parcel and the rental share are determined by bargaining. To solve this new bargaining problem we maximize $N$ in (11) with respect to $r$ and $h$, subject to the land availability constraint $h + (n-1)\bar{h} \leq \alpha$, and taking into account the fact that tenants are maximizing on their own account: $(1-r)F_2 = w$ and $(1-r)F_1 \geq 0$. The conditions for a Nash solution (maximum $N$) are obtained by setting

$$(14) \quad \frac{\partial N}{\partial r} = 0$$

$$(15) \quad \left.\begin{aligned} &\frac{\partial N}{\partial h} \geq 0 \\ &h + (n-1)\bar{h} \leq \alpha \end{aligned}\right\} \text{complementarily}$$

Suppose the land availability constraint does not bind and $(1-r)F_1 > 0$. We have, from (15),

$$(16) \quad (1-r)F_1 \cdot F(h, l_1) + \left[F_1 + F_2 \cdot \frac{dl_1}{dh}\right] \cdot [(1-r)F(h, l_1) - wl_1)] = 0$$

Tenants' employment decisions imply $dl_1/dh = -F_{21}/F_{22} > 0$ for well-behaved pro-

duction functions. Also, $(1-r)F(h, l_1) - wl_1 > 0$ if production is at all undertaken. But then condition (16) is not satisfied, for all terms on the left-hand side are positive. Hence, $F_1 = 0$ if the land availability constraint is not binding. By symmetry, the landlord must reach the same contract with all tenants, so that $h = h^0 = \bar{h}$ and $r = r^0 = \bar{r}$, with $F_1[h^0(r^0), l^0(r^0)] = 0$. We are then back to the first case. Alternatively, suppose that the land availability constraint binds; but then we are in our original formulation of the problem, in Section II.

We have thus established that if any land lies uncultivated, then the marginal productivity of land is zero and the rental share is positive. This result contrasts both with the competitive solution, in which the rental share is zero if $F_1 = 0$, and with the monopoly solution, in which both the rental share and $F_1$ are positive in the face of land lying uncultivated.

## IV. Concluding Remarks

In this paper, we have used bargaining theory to derive determinate solutions for the rental share. We have adopted this approach because both the nature of the contract and the features of the agrarian system necessitate it, while competitive formulations must ignore such considerations. In practice, of course, the bargains which are struck need not conform exactly to those predicted by Nash's solution. But if Nash's solution is empirically valid, and given that the bargaining process is repeated year after year in circumstances of unchanging technology and social institutions, then there will be a long-term tendency for contracts to center on some average value. Moreover, if in most situations the solution is not very different from one-half, agents will save themselves the bother of detailed arithmetic and settle for that magic number, which has also the advantage of sounding equitable. Thus that particular rental share is elevated to the status of a social norm.

It goes without saying that some of the drastic simplifications in our formulation may vitiate the analysis. In particular, when a would-be tenant can deal with more than one landlord his bargaining position is stronger than in our formulation. But given the spatial dimensions of agricultural production and the resulting costs of cultivating widely scattered parcels, our neglect of the alternative opportunities stemming from the existence of other landlords may not prove to be too serious. To introduce such considerations into the analysis would entail very considerable complications, which we have chosen not to pursue at this stage.

## REFERENCES

**P. K. Bardhan and T. N. Srinivasan**, "Cropsharing Tenancy in Agriculture: A Theoretical and Empirical Analysis," *Amer. Econ. Rev.*, Mar. 1971, *61*, 48–64.

——— **and** ———, "Cropsharing Tenancy in Agriculture: Rejoinder," *Amer. Econ. Rev.*, Dec. 1974, *64*, 1067–69.

**C. Bell**, "Some Tests of Alternative Theories of Sharecropping Using Evidence from Northeast India," *J. Develop. Stud.*, forthcoming.

**K. Bharadwaj**, *Production Conditions in Indian Agriculture*, London 1974.

**S. K. Chakraverty**, *Problems of Small Farmers of Kosi Area*, Patna 1969.

**S. N. S. Cheung**, *The Theory of Share Tenancy*, Chicago 1969.

**R. H. Day**, "The Economics of Technological Change and the Demise of the Sharecropper," *Amer. Econ. Rev.*, June 1967, *57*, 427–49.

**W. D. Hopper**, "Allocation Efficiency in a Traditional Indian Agriculture," *J. Farm Econ.*, Aug. 1965, *43*, 611–24.

**R. D. Luce and H. Raiffa**, *Games and Decisions*, New York 1957.

**A. E. Marshall**, *Principles of Economics*, London 1920.

**J. Nash**, "Two Person Cooperative Games," *Econometrica*, Jan. 1953, *21*, 128–40.

**D. M. G. Newbery**, "The Choice of Rental Contract in Peasant Agriculture," in L. Rey-

# Consumer's Surplus Without Apology

By ROBERT D. WILLIG*

The purpose of this paper is to settle the controversy surrounding consumer's surplus[1] and, by so doing, to validate its use as a tool of welfare economics. I will show that observed consumer's surplus can be rigorously utilized to estimate the unobservable compensating and equivalent variations—the correct theoretical measures of the welfare impact of changes in prices and income on an individual.

I derive precise upper and lower bounds on the percentage errors of approximating the compensating and equivalent variations with consumer's surplus. These bounds can be explicitly calculated from observable demand data, and it is clear that in most applications the error of approximation will be very small. In fact, the error will often be overshadowed by the errors involved in estimating the demand curve. The results in no way depend upon arguments about the constancy of the marginal utility of income.

Consequently, this paper supplies specific empirical criteria which can replace the apologetic caveats frequently employed by those who presently apply consumer's surplus. Moreover, the results imply that consumer's surplus is usually a very good approximation to the appropriate welfare measures.

To preview, below I establish the validity of these rules of thumb: For a single[2] price change, if $|\bar{\eta}A/2m^0| \leq .05$, $|\underline{\eta}A/2m^0| \leq .05$, and if $|A/m^0| \leq .9$, then

$$\frac{\underline{\eta}\,|A|}{2m^0} \leq \frac{C-A}{|A|} \leq \frac{\bar{\eta}\,|A|}{2m^0} \tag{1}$$

and

$$\frac{\underline{\eta}\,|A|}{2m^0} \leq \frac{A-E}{|A|} \leq \frac{\bar{\eta}\,|A|}{2m^0} \tag{2}$$

Here, $A$ = consumer's surplus area under the demand curve and between the two prices (positive for a price increase and negative for a price decrease)

$C$ = compensating variation corresponding to the price change

$E$ = equivalent variation corresponding to the price change

$m^0$ = consumer's base income

$\bar{\eta}$ and $\underline{\eta}$ = respectively the largest and smallest values of the income

* Economic Research Department, Bell Laboratories. This paper is drawn from doctoral research done at Stanford University under the guidance of Jim Rosse. I am grateful for his support and for the standards of professional excellence he tried to teach me. I would also like to thank Megina Jack for considerable editorial assistance.

[1] Throughout, the term consumer's surplus is used to refer to the area to the left of an individual's fixed-income (Marshallian) demand curve and between the relevant price horizontals. The concept of consumer's surplus originated in 1844 (see Jules Dupuit) and has been controversial ever since. Alfred Marshall, who popularized the tool, stipulated that for it to be validly used the marginal utility of money must be constant (Marshall, p. 842 or David Katzner, p. 152). However, Harold Hotelling wrote that consumer's surpluses "give a meaningful measure of social value. This breaks down if the variations under consideration are too large a part of the total economy of the person . . . " (p. 289). John Hicks too, stated only a gentle caution: "In order that the Marshallian measure of consumer's surplus should be a good measure, one thing alone is needful—that the income effect should be small" (p. 177). More recently, though, Paul Samuelson (pp. 194–95) concluded that consumer's surplus is a worse than useless concept (because it confuses), and I.M.D. Little (p. 180) agreed, calling it no more than a "theoretical toy." Nonetheless, theorists and cost-benefit analysts have persisted in their use of the tool. For justification they resort (see E. J. Mishan, pp. 337–38, for example), with no formal theoretical support, to statements similar to those quoted above from Hotelling and Hicks.

[2] While I restrict attention to single price changes here, analogous, but more complex formulae are derived for multiple price changes in my papers (1973a, b).

elasticity of demand in the region under consideration.

The formulae place observable bounds on the percentage errors of approximating the $C$ or $E$ conceptual measures with observable $A$. For example, if the consumer's measured income elasticity of demand is 0.8 and if the surplus area under the demand curve between the old and new prices is 5 percent of income, then the compensating variation is within 2 percent of the measured consumer's surplus.

The ratio $|A|/m^0$ can be interpreted as a measure of the proportional change in real income due to the price change.[3] In most applications, the ratio will be very small. Measured income elasticities of demand tend to cluster closely about 1.0, with only rare outliers. Thus it can be expected that $\bar{\eta}|A|/2m^0$, the most important of the terms in (1) and (2), will usually be small enough to permit conscious and unapologetic substitution of $A$ for $C$ or $E$ in studies of individual welfare.[4]

Should $\underline{\eta}|A|/2m^0$ be large, $A$ would not be close to $C$ and $E$. For such rare cases, formulae are provided below in Section IV which enable the estimation of $C$ and $E$ from the observable $\bar{\eta}$, $\underline{\eta}$, $m^0$, and $A$.

## I. The Compensating and Equivalent Variations

In this section, I present definitions of conceptual tools to measure the costs or benefits of price changes to an individual consumer. While these theoretical measures are not directly observable, the analysis that follows in succeeding sections will show that they can be empirically estimated with consumer's surplus.

Throughout I will be assuming that the consumer behaves as though he were choosing his consumption bundle $X=X^1, X^2, \ldots, X^n$ to maximize an increasing strictly quasi-concave ordinal utility function $U(X)$ subject to the budget constraint $\sum p_i X^i = m$. The resulting demand functions, denoted $X^i(p, m)$, are assumed to be differentiable. The indirect utility function, defined by

$$l(p, m) \equiv U[X^1(p, m), X^2(p, m), \ldots, X^n(p, m)]$$

relates the price and income parameters to the maximum level of utility the consumer can achieve under the resulting budget constraint. Clearly, by nonsatiation, $l(p, m)$ is monotone increasing in income $m$, and decreasing in prices $p$.

The indirect utility function can be used to make statements about individual welfare. Let the base, initial situation be characterized by prices $p^0$ and income $m^0$, while an alternative situation can be summarized by $p', m'$. The economic well-being of the consumer in the different situations can be compared by means of the ordinal ranking of the numbers $l(p^0, m^0)$ and $l(p', m')$.

Another way to effect this welfare test is to compare the income change $m'-m^0$ to the smallest income adjustment needed to make the consumer indifferent to the change in prices from $p^0$ to $p'$. If $m'-m^0$ is larger, then welfare is greater in the new situation, and inversely.

This test level of income adjustment is called the compensating income variation, denoted by $C$ below. Symbolically,

$$(3) \qquad l(p^0, m^0) = l(p', m^0 + C)$$

The welfare test above

$$(4) \quad l(p', m') \gtreqless l(p^0, m^0) \quad \text{as} \quad m' - m^0 \gtreqless C$$

follows immediately from (3) by nonsatiation. Thus the compensating variation is an individual's cost-benefit concept which makes price changes perfectly commensurable with changes in

[3] Or the ratio can be interpreted using the words of Hotelling quoted in fn. 1 as the relative size of the variation.

[4] Formulae (1) and (2) reflect the cautions (see fn. 1) of both Hotelling and Hicks.

income.

Similarly, the equivalent variation in income ($E$) can be defined[5] by

$$(5)\qquad l(p^0, m^0 - E) = l(p', m^0)$$

In words, $-E$ is the income change which has the same welfare impact on the consumer in the base situation as have the changes in prices from $p^0$ to $p'$. It reduces the impacts of different price changes down to the single dimension of income. As such, the equivalent variation concept can be used to rank the consumer's levels of well-being under various sets of prices. With the definitions $l(p^0, m^0-E')=l(p', m^0)$ and $l(p^0, m^0-E'')=l(p'', m^0)$, these welfare tests, too, follow from nonsatiation:

$$(6)\qquad \begin{aligned} l(p', m^0) &\gtreqless l(p'', m^0) \quad \text{as} \quad E'' \gtreqless E' \\ l(p', m^0) &\gtreqless l(p^0, m') \quad \text{as} \quad m^0 - E \gtreqless m' \end{aligned}$$

The welfare tests (4) and (6) show that the compensating and equivalent variations are cost-benefit concepts which can be used to evaluate the impact of microeconomic policy on an individual.[6] These concepts derive practical importance from the fact that they can be estimated from observable consumer's surplus.

## II. Consumer's Surplus

The compensating and equivalent variations can be most incisively studied and related to consumer's surplus by means of the income compensation function.[7] This is denoted by $\mu(p|p^0, m^0)$ and is defined to be the least income required by the consumer when he faces prices $p$ to achieve the same utility level he could enjoy (by maximizing behavior) under the parameters $p^0$, $m^0$. Thus, by definition,

$$(7)\qquad l[p, \mu(p|p^0, m^0)] = l(p^0, m^0)$$

Trivially, we have

$$(8)\qquad \mu(p^0|p^0, m^0) = m^0$$

Now, we can see that the compensating and equivalent variations can be expressed or redefined in terms of the income compensation function. From (3), $m^0+C=\mu(p'|p^0, m^0)$, or combining with (8),

$$(9)\qquad C = \mu(p'|p^0, m^0) - \mu(p^0|p^0, m^0)$$

Similarly, from (5), $m^0-E=\mu(p^0|p', m^0)$, or

$$(10)\qquad E = \mu(p'|p', m^0) - \mu(p^0|p', m^0)$$

These relationships serve as the bridge to consumer's surplus.

It is well known[8] that

$$(11)\qquad \frac{\partial\mu(p|p^0, m^0)}{\partial p_i} = X^i(p, \mu(p|p^0, m^0))$$

This system of partial differential equations, together with the boundary condition (8), is the heart of analytical welfare economics.[9] The compensating and equivalent variations, or any measure of individual welfare that accepts the individual's own consumption preferences, can be calculated from the complete demand functions via (11) and (8).

Restricting attention to changes in a single price, $p_1$, let $p^0=(p_1^0, p_2^0, \ldots, p_n^0)$ and $p'=(p_1', p_2^0, \ldots, p_n^0)$. Use the Fundamental Theorem of Calculus and (11) to rewrite (9) and (10) as

[5] The definitions (3) and (5) correspond to those of Hicks, p. 177, and Samuelson, p. 199.

[6] They also can serve as building blocks for methodologies to make social welfare judgments. The Compensation Principle is a well-known example (see Tibor Scitovsky).

[7] This theoretical tool was introduced by Lionel McKenzie, and definitively studied by Leonid Hurwicz and Hirofumi Uzawa.

[8] See Hurwicz and Uzawa for a state-of-the-art derivation. Heuristically, (11) says that the first-order income change, $d\mu$, required to compensate for the price increase, $dp_1$, is just the augmentation needed to buy the old consumption bundle, $X(p, \mu,(p|p^0, m^0))$, at the new prices $p_1+dp_1, p_2, \ldots, p_n$, rather than at the old prices $p$. The irrelevance to this calculation of the concomitant substitution effects is the result of the envelope theorem.

[9] This point of view was taken by Herbert Mohring.

$$(12)\quad C = \int_{p_1^0}^{p_1'} X^1(p_1, \overset{0}{p_2}, \ldots, \overset{0}{p_n}, \mu(p_1, \overset{0}{p_2}, \ldots, \overset{0}{p_n} \,|\, \overset{0}{p_1}, \overset{0}{p_2}, \ldots, \overset{0}{p_n}, \overset{0}{m}))dp_1$$

$$(13)\quad E = \int_{p_1^0}^{p_1'} X^1(p_1, \overset{0}{p_2}, \ldots, \overset{0}{p_n}, \mu(p_1, \overset{0}{p_2}, \ldots, \overset{0}{p_n} \,|\, p_1', \overset{0}{p_2}, \ldots, \overset{0}{p_n}, \overset{0}{m}))dp_1$$

These formulae express the compensating and equivalent variations as areas under demand curves, between the old and new price horizontals. The demand curves are not Marshallian in that the income parameters are not constant. Instead, they are Hicksian compensated demand curves, because the income parameters include compensation which varies with the price to keep the consumer at a constant level of utility. The only distinction between $C$ and $E$ in (12) and (13) is the level of utility the compensation is designed to reach.

Referring to Figure 1, $C$ is the area $p_1^0 p_1' be$ under the demand curve compensated to $l(p^0, m^0)$. This curve crosses the Marshallian curve $X^1(p, m^0)$ at $p_1^0$, since $\mu(p^0 | p^0, m^0) = m^0$. With $p_1' > p_1^0$, if $X^1$ is noninferior ($\partial X^1/\partial m \geq 0$) this compensated curve lies above the Marshallian one for $p_1 > p_1^0$, since $\mu(p_1, p_2^0, \ldots, p_n^2 | p^0, m^0) \geq m^0$ whenever $p_1 > p_1^0$. Similarly, $E$ is the area $p_1^0 p_1' af$ under the demand curve compensated to $l(p', m^0)$. This Hicksian curve crosses the Marshallian one at $p_1'$, and lies below it for $p_1 < p_1'$. The area usually called consumer's surplus is $p_1^0 p_1' ae$, defined by the observable Marshallian demand curve. Denoting this area by $A$, we have, then, $C \geq A \geq E$, for noninferior $X^1$ (the inequalities reverse for $X^1$ inferior). Of course, it also follows immediately that if there is no income effect ($\partial X^1/\partial m \equiv 0$), $C = A = E$.

These qualitative results may be useful for some cost-benefit analyses. For example, suppose a policy would raise both an individual's income and the price of a noninferior good. If the observable con-

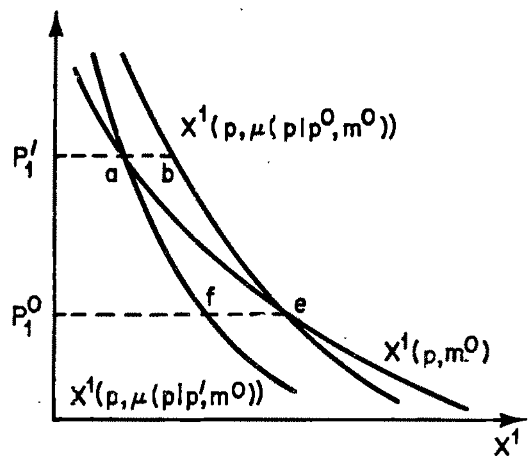



FIGURE 1

sumer's surplus area $A$ were greater than the income boost, it could be inferred from the inequality that $C$ also would be greater. Then, from the welfare test (4), an analyst could conclude that the policy would be injurious to the consumer.

However, usually more information than this is needed about $C$ and $E$. What is required is a methodology to estimate the welfare measures from observable data. In the next section I show how $C$ and $E$ can be explicitly calculated from observables when the income elasticity of demand is constant.

## III. Constant Income Elasticity

Constant income elasticity of demand for $X^1$ means that

$$\frac{\partial X^1(p, m)}{\partial m} \frac{m}{X^1(p, m)} \equiv \eta$$

Then, we have the simple differential equation $dX^1/X^1 = \eta(dm/m)$ which can be integrated from $X^1(p, m^0)$ to yield

$$X^1(p, m) = X^1(p, m^0)\left[\frac{m}{m^0}\right]^\eta$$

The entire income compensation function can be derived by substituting this expression into (11) and solving the result-

ing differential equation with boundary condition (8). We have, suppressing unchanging arguments,

$$\frac{d\mu}{dp_1} = X^1(p_1, \mu) = X^1(p_1, m^0)\left[\frac{\mu}{m^0}\right]^\eta$$

or

$$\mu^{-\eta} d\mu = (m^0)^{-\eta} X^1(p_1, m^0)dp_1$$

Then, integration between $p_1^0$ and $p_1'$, remembering that $\mu(p_1^0) = m^0$, yields

$$(14)\quad \frac{[\mu(p_1')]^{1-\eta} - [m^0]^{1-\eta}}{1-\eta} = (m^0)^{-\eta}\int_{p_1^0}^{p_1'} X^1(p_1, m^0)dp_1$$

for $\eta \neq 1$, and for $\eta = 1$,

$$ln\,\mu(p_1') - ln\, m^0 = \frac{1}{m^0}\int_{p_1^0}^{p_1'} X^1(p_1, m^0)dp_1$$

Hence, after rearranging we have these explicit expressions for the income compensation function:

$$(15)\quad \mu(p_1' \mid p_1^0, m^0) = m^0\left[1 + \left(\frac{1-\eta}{m^0}\right)\int_{p_1^0}^{p_1'} X^1(p_1, m^0)dp_1\right]^{1/1-\eta}$$

$\eta \neq 1$

$$(16)\quad \mu(p_1' \mid p_1^0, m^0) = (m^0)\exp\left[\frac{1}{m^0}\int_{p_1^0}^{p_1'} X^1(p_1, m^0)dp_1\right]$$

$\eta = 1$

These give the welfare measure $\mu$ in terms of the potentially observable constant income elasticity of demand and the consumer's surplus area under the Marshallian demand curve. Let us denote this area by

$$(17)\quad A \equiv \int_{p_1^0}^{p_1'} X^1(p_1, m^0)dp_1$$

From (15), we see that if $\eta = 0$, $\mu(p_1' \mid p_1^0, m^0) = m^0 + A$. However, from (16) we see that if preferences are homothetic, the consequent unitary $\eta$ does not imply any equalities among $C$, $E$, and $A$. Below, for expositional convenience, I ignore the case $\eta = 1$.

Recalling the definitions of $C$ and $E$, (9) and (10), and loosely applying to (15) this Taylor approximation,

$$(1+t)^{1/1-\eta} \approx 1 + \frac{t}{1-\eta} + \frac{\eta t^2}{2(1-\eta)^2}$$

(where $\approx$ means "approximately equal to"), we get:

$$C \approx A + \frac{\eta A^2}{2m^0}, \qquad E \approx A - \frac{\eta A^2}{2m^0}$$

$$\frac{C-A}{A} \approx \frac{\eta A}{2m^0} \quad \text{and} \quad \frac{A-E}{A} \approx \frac{\eta A}{2m^0}$$

This was the striking result on the percentage error of approximating $C$ with $A$ which was previewed in the introduction. The next section will establish this formula rigorously for nonconstant income elasticity of demand.

## IV. Estimation Results

Assume that in the region of price-income space under consideration,[10] $\bar{\eta}$ and $\underline{\eta}$ are upper and lower bounds, respectively, on $(\partial X^1(p, m)/\partial m)(m/X^1(p, m))$, with neither equal to 1.[11] It follows from the Mean Value Theorem that

$$(18)\quad \left(\frac{m_2}{m_1}\right)^{\underline{\eta}} \leq \frac{X^1(p, m_2)}{X^1(p, m_1)} \leq \left(\frac{m_2}{m_1}\right)^{\bar{\eta}} \quad \text{for } m_2 \geq m_1$$

Let us consider the welfare impact of a price increase from $p_1^0$ to $p_1'$. Since $\mu(p_1 \mid p^0, m^0) \geq \mu(p_1^0 \mid p^0, m^0)$ for $p_1 \geq p_1^0$, we can set $m_2 = \mu(p_1)$ and $m_1 = \mu(p_1^0) = m^0$ in (18):

[10] This region is $\{(p, m): p_1 = \alpha p_1^0 + (1-\alpha)p_1', 0 \leq \alpha \leq 1; p_i = p_i^0, i \neq 1; m = \gamma m^0 + (1-\gamma)\mu(p \mid p^0, m^0), 0 \leq \gamma \leq 1;$ and $X^1(p, m) > 0\}$.

[11] Either $\bar{\eta}$ or $\underline{\eta}$ can be arbitrarily close to 1.

$$\left[\frac{\mu(p_1)}{m^0}\right]^{\underline{\eta}} \leq \frac{X^1(p_1, \mu(p_1))}{X^1(p_1, m^0)} \leq \left[\frac{\mu(p_1)}{m^0}\right]^{\bar{\eta}}$$

Rearranging, and substituting from (11) yields

$$0 \leq X^1(p, m^0)^{-\underline{\eta}} \leq \frac{\partial\mu(p)}{\partial p_1}[\mu(p)]^{-\underline{\eta}} = \partial\left[\frac{\mu(p)^{1-\underline{\eta}}}{1-\underline{\eta}}\right] \Big/ \partial p_1$$

and

$$0 \leq \partial\left[\frac{\mu(p)^{1-\bar{\eta}}}{1-\bar{\eta}}\right] \Big/ \partial p_1 = \frac{\partial\mu(p)}{\partial p_1}\mu(p)^{-\bar{\eta}} \leq X^1(p, m^0)(m^0)^{-\bar{\eta}}$$

Integrating these relationships with respect to $p_1$ between $p_1^0$ and $p_1'$ (as in (14)) preserves the inequalities. Rearrangement of the resulting relationships yields these bounds:

$$(19)\quad m^0\left[1 + (1-\underline{\eta})\frac{A}{m^0}\right]^{1/1-\underline{\eta}} \leq \mu(p' \mid p^0, m^0) \leq m^0\left[1 + (1-\bar{\eta})\frac{A}{m^0}\right]^{1/1-\bar{\eta}}$$

provided $\underline{\eta}, \bar{\eta} \neq 1,\ 1 + (1-\underline{\eta})\frac{A}{m^0} > 0$

and $1 + (1-\bar{\eta})\frac{A}{m^0} > 0$

For the case of a price decrease from $p_1^0$ to $p_1'$, since $\mu(p_1 \mid p_1^0, m^0) \leq m^0$ for $p_1 \leq p_1^0$, we can set $m_2 = m^0$ and $m_1 = \mu(p_1)$ in (18), and then follow the same sequence of steps. Once again, (19) emerges, but reference to (17) shows that here $A$ is negative.

Invoking the definition (9), (19) can be rewritten as

$$(20)\quad \frac{\left[1 + (1-\underline{\eta})\frac{A}{m^0}\right]^{1/1-\underline{\eta}} - 1 - \frac{A}{m^0}}{|A|/m^0} \leq \frac{C-A}{|A|} \leq \frac{\left[1 + (1-\bar{\eta})\frac{A}{m^0}\right]^{1/1-\bar{\eta}} - 1 - \frac{A}{m^0}}{|A|/m^0}$$

Also, using (10) and reversing the roles of $p'$ and $p^0$ in (19) (but not in the definition of $A$) gives

$$(21)\quad \frac{\left[1 - (1-\underline{\eta})\frac{A}{m^0}\right]^{1/1-\underline{\eta}} - 1 + \frac{A}{m^0}}{|A|/m^0} \leq \frac{A-E}{|A|} \leq \frac{\left[1 - (1-\bar{\eta})\frac{A}{m^0}\right]^{1/1-\bar{\eta}} - 1 + \frac{A}{m^0}}{|A|/m^0}$$

The measures of a consumer's welfare can be tightly estimated from observables via (19)–(21), regardless of the size of $A/m^0$, if $1 \pm (1-\underline{\eta})A/m^0 > 0$, $1 \pm (1-\bar{\eta})A/m^0 > 0$, and if $\underline{\eta}$, and $\bar{\eta}$ are sufficiently close in value.[12] Of course, in the limit, as $\underline{\eta}$ approaches $\bar{\eta}$, (19) reduces to the constant elasticity formula (15). Moreover, we shall see that if the absolute values of $\underline{\eta}A/2m^0$ and $\bar{\eta}A/2m^0$ are small, then (20) and (21) reduce to elegant rules of thumb.

Table 1 displays the numerical values of the following coefficients for selected choices of $\eta$ and $a$:

[12] The most plausible cause of the negation of these conditions is $(\partial X^1/\partial m)(m/X^1) \to \infty$. However, regions in which $X^1$ is identically zero can be ignored, since there both $\mu$ and $A$ are unchanging. To handle the case in which $X^1 = 0$ and $\partial X^1/\partial m \neq 0$ near the boundary of the relevant region, bounds on $\mu$ can be derived from bounds on $\partial X^1/\partial m$. Because these are generally more gross than (19), the best approach is to take this tack only in the vicinity of the singularity, use (19) on the rest of the path of integration, and splice the sets of inequalities together. The formulae for such procedures can be found in my 1973a, b papers. An explicit solution for $\mu$ when $\partial X^1/\partial m$ is independent of $m$ is also reported there.

TABLE 1[a]

| η \ a | .001 | .005 | .010 | .020 | .030 | .040 | .050 | .075 | .100 | .150 | .200 | .250 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -2.00 | -.001 | -.005 | -.010 | -.020 | -.030 | -.040 | -.050 | -.075 | -.100 | -.150 | -.200 | -.250 |
| | -.001 | -.005 | -.010 | -.019 | -.029 | -.038 | -.046 | -.067 | -.086 | -.121 | -.152 | -.180 |
| | -.001 | -.005 | -.010 | -.021 | -.032 | -.043 | -.054 | -.086 | -.121 | -.205 | -.316 | -.480 |
| -1.01 | -.001 | -.003 | -.005 | -.010 | -.015 | -.020 | -.025 | -.038 | -.051 | -.076 | -.101 | -.126 |
| | -.001 | -.003 | -.005 | -.010 | -.015 | -.019 | -.024 | -.035 | -.046 | -.066 | -.085 | -.102 |
| | -.001 | -.003 | -.005 | -.010 | -.016 | -.021 | -.027 | -.041 | -.056 | -.090 | -.129 | -.174 |
| .30 | .000 | .001 | .002 | .003 | .005 | .006 | .008 | .011 | .015 | .023 | .030 | .038 |
| | .000 | .001 | .002 | .003 | .005 | .006 | .008 | .011 | .015 | .023 | .029 | .036 |
| | .000 | .001 | .002 | .003 | .005 | .006 | .008 | .011 | .015 | .023 | .031 | .039 |
| .50 | .000 | .001 | .003 | .005 | .008 | .010 | .013 | .019 | .025 | .038 | .050 | .063 |
| | .000 | .001 | .003 | .005 | .008 | .010 | .013 | .019 | .025 | .038 | .050 | .063 |
| | .000 | .001 | .003 | .005 | .008 | .010 | .013 | .019 | .025 | .038 | .050 | .063 |
| .70 | .000 | .002 | .004 | .007 | .011 | .014 | .018 | .026 | .035 | .053 | .070 | .088 |
| | .000 | .002 | .004 | .007 | .011 | .014 | .018 | .027 | .035 | .054 | .072 | .090 |
| | .000 | .002 | .004 | .007 | .010 | .014 | .017 | .026 | .035 | .051 | .068 | .085 |
| .90 | .000 | .002 | .005 | .009 | .014 | .018 | .023 | .034 | .045 | .068 | .090 | .113 |
| | .000 | .002 | .005 | .009 | .014 | .018 | .023 | .034 | .046 | .070 | .095 | .120 |
| | .000 | .002 | .004 | .009 | .013 | .018 | .022 | .033 | .044 | .065 | .085 | .105 |
| 1.01 | .001 | .003 | .005 | .010 | .015 | .020 | .025 | .038 | .051 | .076 | .101 | .126 |
| | .001 | .003 | .005 | .010 | .015 | .020 | .026 | .039 | .052 | .080 | .108 | .138 |
| | .001 | .003 | .005 | .010 | .015 | .020 | .025 | .037 | .049 | .072 | .094 | .116 |
| 1.10 | .001 | .003 | .006 | .011 | .017 | .022 | .028 | .041 | .055 | .083 | .110 | .138 |
| | .001 | .003 | .006 | .011 | .017 | .022 | .028 | .043 | .057 | .088 | .119 | .152 |
| | .001 | .003 | .006 | .011 | .016 | .022 | .027 | .040 | .053 | .078 | .102 | .125 |
| 1.20 | .001 | .003 | .006 | .012 | .018 | .024 | .030 | .045 | .060 | .090 | .120 | .150 |
| | .001 | .003 | .006 | .012 | .018 | .024 | .031 | .047 | .063 | .097 | .132 | .169 |
| | .001 | .003 | .006 | .012 | .018 | .024 | .029 | .043 | .057 | .084 | .110 | .134 |
| 1.50 | .001 | .004 | .008 | .015 | .023 | .030 | .038 | .056 | .075 | .113 | .150 | .188 |
| | .001 | .004 | .008 | .015 | .023 | .031 | .039 | .059 | .080 | .125 | .173 | .224 |
| | .001 | .004 | .007 | .015 | .022 | .029 | .036 | .054 | .070 | .102 | .132 | .160 |
| 2.00 | .001 | .005 | .010 | .020 | .030 | .040 | .050 | .075 | .100 | .150 | .200 | .250 |
| | .001 | .005 | .010 | .020 | .031 | .042 | .053 | .081 | .111 | .176 | .250 | .333 |
| | .001 | .005 | .010 | .020 | .029 | .038 | .048 | .070 | .091 | .130 | .167 | .200 |
| 3.00 | .002 | .008 | .015 | .030 | .045 | .060 | .075 | .113 | .150 | .225 | .300 | .375 |
| | .002 | .008 | .015 | .031 | .047 | .064 | .082 | .129 | .180 | .302 | .455 | .657 |
| | .002 | .008 | .015 | .029 | .043 | .056 | .069 | .100 | .129 | .180 | .226 | .266 |
| 5.00 | .003 | .013 | .025 | .050 | .075 | .100 | .125 | .188 | .250 | .375 | .500 | .625 |
| | .003 | .013 | .026 | .053 | .082 | .114 | .147 | .244 | .362 | .716 | 1.477 | ** |
| | .002 | .012 | .024 | .047 | .069 | .089 | .109 | .154 | .193 | .261 | .317 | .364 |
| 10.00 | .005 | .025 | .050 | .100 | .150 | .200 | .250 | .375 | .500 | .750 | 1.000 | 1.250 |
| | .005 | .026 | .053 | .115 | .186 | .271 | .374 | .774 | 1.916 | ** | ** | ** |
| | .005 | .024 | .047 | .089 | .126 | .160 | .191 | .257 | .312 | .396 | .460 | .509 |

[a] Each group of three numbers includes, from the top, $\eta a/2$, $[(1+(1-\eta)a)^{1/1-\eta}-1-a]/a$, and $[(1-(1-\eta)a)^{1/1-\eta}-1+a]/a$. The entry ** indicates that $(1+(1-\eta)a)<0$.

$$\frac{\eta a}{2}, \quad \frac{[1+(1-\eta)a]^{1/1-\eta}-1-a}{a}$$

and

$$\frac{[1-(1-\eta)a]^{1/1-\eta}-1+a}{a}$$

The latter two expressions encompass the forms of the bounds in (20) and (21), when $a$ is interpreted as $|A|/m^0$.[13] It can

[13] For example, the value of the lower bound in (20) when $\eta=2$ and $A/m^0=-.05$ is .048. This can be found in Table 1 as the value of $[(1-(1-\eta)a)^{1/1-\eta}-1+a]/|a|$ when $\eta=2$ and $a=.05$.

be readily seen from the table that for the ranges of parameter values studied,[14] when $|\eta a/2|$ is small (say less than .05), $\eta a/2$ is close enough (within .005) to the actual bounds for most practical purposes. This numerical observation corroborates the loose application to (15) of the Taylor Series expansion in Section III. More importantly, it establishes the rules of thumb previewed in (1) and (2).[15]

Addition of (1) and (2) yields a check on the numerical proximity of $C$ and $E$: when $|\underline{\eta}A/2m^0| \leq .05$, $|\bar{\eta}A/2m^0| \leq .05$, and $|A/m^0| \leq .9$,

$$\frac{\underline{\eta}|A|}{m^0} \leq \frac{C-E}{|A|} \leq \frac{\bar{\eta}|A|}{m^0} \tag{23}$$

So, the analysis hinges on the magnitudes of $\eta$ and $A/m^0$. As discussed in the introduction, in most practical applications $|\eta A/2m^0|$ and $|A/m^0|$ are likely to be small enough for the rules of thumb to apply. If not, equations (19)–(21) and Table 1 will be useful. Even if the calculated error bounds are too large to be ignored, the compensating and equivalent variations may still be usefully estimated from the data via the formulae.

## V. Individual Welfare and Consumer's Surplus

With the approximation results in hand, let us return to the question of how to make statements about individual welfare, based on observable data. Remember from (4) that $l(p', m') \gtreqless l(p^0, m^0)$ as $m' - m^0 \gtreqless C$. With the empirical information that $\underline{C} \leq C \leq \bar{C}$, where $\underline{C}$ and $\bar{C}$ can be calculated from (20) or (22), it can be concluded that

$$\begin{aligned} (24)\quad & l(p', m') > l(p^0, m^0), \quad \text{if } m' - m^0 > \bar{C} \\ & l(p', m') < l(p^0, m^0), \quad \text{if } m' - m^0 < \underline{C} \end{aligned}$$

[14] These seem to include most values that would be found for these parameters in actual applications.

[15] When $|\underline{\eta}A/2m^0| \leq .05$ and $|\bar{\eta}A/2m^0| \leq .05$, it suffices for $1 \pm (1-\underline{\eta})A/m^0 > 0$ and $1 \pm (1-\bar{\eta})A/m^0 > 0$ that $|A/m^0| < .9$.

If $\underline{C}$ and $\bar{C}$ are close in value, (24) provides a welfare test of considerable power.[16] If $|\bar{\eta}A/2m^0|$ and $|\underline{\eta}A/2m^0|$ are small enough, both $\underline{C}$ and $\bar{C}$ can be safely replaced in (24) by $A$. Otherwise, they can be calculated from $\underline{\eta}$, $\bar{\eta}$, $A$, and $m^0$.

To conclude, at the level of the individual consumer, cost-benefit welfare analysis can be performed rigorously and unapologetically by means of consumer's surplus.

[16] Another welfare comparison (which may be useful for an analysis of social welfare with a Bergsonian social welfare function) is made possible by the fact (see Hurwicz and Uzawa) that $\mu(p^0|p, m)$, viewed as a function of $p$ and $m$, is a proper indirect utility function.

$$\mu(p^0|p, m) = E + m$$

where $E$ is the equivalent variation associated with a change from $p^0$ to $p$. Hence this particular ordinal indirect utility function can be exactly expressed by areas under compensated demand curves, as in (13), or it can be estimated from consumer's surplus via (19), (21), or (2).

## REFERENCES


**K. J. Arrow and T. Scitovsky,** *Readings in Welfare Economics,* vol. 12, Homewood 1969.

**J. Dupuit,** "On the Measurement of the Utility of Public Works," (1844), translated and reprinted in K. J. Arrow and T. Scitovsky, eds., *Readings in Welfare Economics,* vol. 12, Homewood 1969, 255–83.

**J. R. Hicks,** *A Revision of Demand Theory,* London 1956.

**H. Hotelling,** "The General Welfare in Relation to Problems of Taxation and of Railway and Utility Rates," reprinted in K. J. Arrow and T. Scitovsky, eds., *Readings in Welfare Economics,* vol. 12, Homewood 1969.

**L. Hurwicz and H. Uzawa,** "On the Integrability of Demand Functions," in J. S. Chipman et al., eds., *Preferences, Utility, and Demand,* New York 1971, 114–48.

**D. Katzner,** *Static Demand Theory,* New York 1970.

**I. M. D. Little,** *A Critique of Welfare Economics,* London 1957.

**A. Marshall,** *Principles of Economics,* 9th ed., New York 1961.

L. W. McKenzie, "Demand Theory without a Utility Index," *Rev. Econ. Stud.*, June 1957, *24*, 185–89.

E. J. Mishan, *Cost-Benefit Analysis: An Introduction*, New York 1971.

H. Mohring, "Alternative Welfare Gain and Loss Measures," *Western Econ. J.*, Dec. 1971, *9*, 349–68.

P. A. Samuelson, *Foundations of Economic Analysis*, Cambridge 1947.

T. Scitovsky, "A Note on Welfare Propositions in Economics," reprinted in K. J. Arrow and T. Scitovsky, eds., *Readings in Welfare Economics*, vol. 12, Homewood 1969.

R. Willig, (1973a) "Consumer's Surplus: A Rigorous Cookbook," tech. rep. no. 98, Economics Series, Inst. for Mathemat. Stud. in the Soc. Sci., Stanford Univ. 1973.

———, (1973b) "Welfare Analysis of Policies Affecting Prices and Products," memo. no. 153, Center for Research in Econ. Growth, Stanford Univ. 1973.

# Trade and Domestic Effects of the Offshore Assembly Provision in the U.S. Tariff

By J. M. FINGER*

The *U.S.* tariff contains an offshore assembly provision, (*OAP*). The tariff on certain imported goods is charged on the value of such goods *less* the value of incorporated domestically produced components, i.e., the tariff on an assembled product is applied only to the value-added in foreign assembly or processing if domestically produced components are used. Such provisions were included in the 1930 *U.S.* tariff,[1] and they are the subject of continuing controversy. Clauses to repeal the offshore assembly provisions[2] in the *U.S.* tariff were included in the recent Mills and Burke-Hartke trade bills (which were not passed by the Congress), and an unsuccessful attempt was made to include such a repeal clause in the recently passed Trade Reform Act of 1974. The AFL-CIO is actively pursuing the repeal of these provisions.

With an *OAP*, the tariff on an imported good is applied only to the value-added in assembly, *if* domestic components are used. One of its effects then is to reduce the effective rate of protection of the assembly activity, and hence to shift demand from domestic to foreign assemblers. Opponents of the provisions stress the associated displacement of domestic assembly activity and the accompanying negative balance-of-trade impact.[3] They frequently cite the rapid rate of increase of *OAP* imports from developing countries (*LDC*); 60 percent per year over the period 1966–72 as compared with 12 percent for other manufactured imports from *LDC*.[4]

But this "tariff break" is available only if domestic components are used—implying a simultaneous shift of demand from foreign to domestic components. Defenders of the *OAP* stress the associated positive effects on domestic activity and on the balance of payments.

The object of this paper is to estimate the static impact of the *OAP* on domestic economic activity (in assembly and in production of components) and on the balance of trade. Estimates of the effect on domestic economic welfare of the *OAP* will also be presented. The effect on investment flows will be taken into account indirectly by the supply parameters of the model, but will not be explicitly estimated. To measure balance of payments rather than simply the balance-of-trade impact, an

* U.S. Treasury Department.

[1] A history of such provisions is included in the Tariff Commission report.

[2] Item 806.30 of the *U.S.* tariff allows such tariff treatment on metal articles exported by *U.S.* firms for processing abroad by a subsidiary or foreign subcontractor, and returned for further processing by the *U.S.* firm. Item 807.00 allows such tariff treatment on any imported article which contains *U.S.* components, regardless of whether the foreign operation is performed by a *U.S.* subsidiary, a foreign subcontractor, or a foreign firm which purchases *U.S.* components for inclusion in products exported to the United States. Over 90 percent of *OAP* imports enter under item 807.00. There are limits to the sort of processing allowed abroad. For details, see the Tariff Commission report, pp. 9 ff.

[3] Throughout, the "balance-of-payments impact" is understood to be a pressure on the exchange rate under a regime of flexible rates.

[4] For imports from all countries, comparable rates of increase are 24 and 17 percent. *OAP* imports from all countries were \$3.4 billion in 1972—about 9 percent of manufactured imports; *OAP* imports from *LDC* were \$1 billion—22 percent of manufactured imports.

estimate of the effect of *OAP* trade on the profit remittances of foreign subsidiaries would also be required.

In Sections II, III, and IV a comparative statics theory of the incidence of *OAP* is formulated, and expressions for the various effects of their repeal are derived. In Section V the effects are estimated by substituting parameter values into the formulae.

## I. The Model and Assumptions

The model is essentially the fixed coefficients, partial equilibrium, effective protection model with less than perfectly elastic import supplies. In order that the shifts from domestic to foreign assembly and from foreign to domestic components may be analyzed, it must incorporate foreign and domestic assembly and foreign and domestic production of components. It thus consists of five equations: the demand for the finished good,[5] supply functions of foreign and of domestic components, and supply functions of foreign and domestic assembly. The model is completed by aggregating across countries to determine the total supplies of assembly and of components, which are then aggregated vertically to determine the supply of the finished good. The intersection of this curve with the demand function defines market equilibrium.

In analyzing the *OAP* we are looking at final goods which are imported by the United States and contain components exported by the United States. In the language of the model, this makes the United States an exporter of components and an importer of assembly.

Operationally, the distinction between "assembly" and "components" is based on the separation of the dutiable from the duty free value of a finished good imported under the *OAP*. This distinction is then used to specify what is defined as assembly and as components in import replacements (domestic assembly of domestic components) and in non-*OAP* imports of the same final good (foreign assembly of foreign components).[6] *OAP* and non-*OAP* imports and domestic import replacements are assumed perfect substitutes for each other.

For example, suppose shirts imported under *OAP* are made from parts which are cut in the United States from *U.S.* fabric. The parts are then shipped outside of the United States, where they are sewn with foreign thread into shirts. Upon entry into the United States, the tariff is charged on the "final" value of the shirts, less the value of the *U.S.*-made parts. Thus the dutiable value includes foreign-made components as well as assembly cost, in the strict sense of the term. Components therefore refers to the shirt parts and assembly to the thread and the sewing. Because we assume fixed coefficients, and because the same tariff rate (the rate for the finished good) is applied to all segments of the foreign value, it is not necessary to distinguish in the model the thread from the sewing. Imported shirts sewn from foreign cloth (and thus ineligible for *OAP* treatment) contain foreign components (the parts) and foreign assembly (the thread and the sewing). We distinguish in the same way the components from the assembly of domestically produced import replacement shirts.

The "quantity of assembly" is defined as the quantity of finished goods produced (assembled), and the "supply of assembly" relates this quantity to the rate of payment for the assembly process, i.e., to the "piece rate" received by a contract assembler or the difference between the price

[5] "Finished" of course is relative to the process being studied. Whether these are consumption, intermediate, or capital goods is irrelevant for our purposes.

[6] The reason for not assuming that all imports of the finished good contain *U.S.* components (i.e., are *OAP* imports) is empirical. For virtually every 7-digit product for which there are *OAP* imports, there are also non-*OAP* imports.

received for finished articles and the price paid for components by an assembler who works for his own account.[7] Supplies of components are supply functions in the usual sense and the demand for the finished good is a demand function in the usual sense.

The foreign supply functions are defined as export functions—the foreign supply of assembly refers to assembly for export, and the foreign supply of components to the foreign supply net of demand for production for the foreign market.[8]

### A. *Notation*

The notation has been made mnemonic as far as possible. The $P$, $Q$, and $V$ represent prices, quantities, and values, while $T$ is the ad valorem tariff rate applicable to the finished good. The subscripts $d$, $f$, and $t$ distinguish domestic, foreign, and total (sum of foreign and domestic), while $a$ and $c$ identify assembly and components. The subscript $j$ refers to the finished good. Elasticities of supply are represented by $\Phi$; $\Gamma$ is the elasticity of domestic demand for the finished good, defined so as to be negative; $C$ is the component's input-output coefficient, evaluated at foreign prices, i.e., $C=P_{fc}/(P_{fa}+P_{fc})$. The $R_{fa}$ and $R_{fc}$ are shares of imports in total assembly and in total production of components for the *U.S.* market. Starred variables refer to proportionate changes (for example, $V^*=dV/V$) for all variables except $T(T^*=dT/(1+T))$.

### B. *Equations of the Model*

In the case in question, in which domestic components ($DC$) are reimported after assembly abroad, the foreign assembler (or a foreign exporter who buys components and has them assembled into goods for export) can choose between $DC$ and foreign components ($FC$).

If there is no $OAP$, the foreign assembler will always choose the components with the lower price, hence at equilibrium

$$P_{dc} = P_{fc} \tag{1}$$

Because of the domestic economy's tariff on $J$,

$$P_{dj} = (1+T)(P_{fa}+P_{fc}) \tag{2}$$

By definition,

$$P_{da} = P_{dj} - P_{dc} \tag{3}$$

and combining the three equations, we have

$$P_{da} = P_{fa} + T(P_{fa}+P_{fc}) \tag{4}$$

If, however, there is an $OAP$, the foreign assembler can avoid a tariff liability of $T\cdot P_{fc}$ by using $DC$. Hence, at equilibrium (1) is replaced by

$$P_{dc} = (1+T)P_{fc} \tag{5}$$

while (2) and (3) remain relevant, so we have

$$P_{da} = (1+T)P_{fa} \tag{6}$$

Thus the effect of removing the $OAP$ is to "shift" the term $T\cdot P_{fc}$ from the difference between the domestic and foreign prices of components to the difference between the domestic and foreign prices of assembly. This can be treated algebraically by introducing the term $X$, which measures the proportion to which the tariff is a simple tariff on imports of $J$ rather than a tariff with an $OAP$. At one extreme, when $X=1$, the tariff is a "pure" tariff on imports of $J$ without an $OAP$, while at the other extreme when $X=0$, the $OAP$ is in effect. We can thus write

$$P_{dc} = (1+(1-X)T)P_{fc} \tag{7}$$

$$P_{da} = (1+T)P_{fa} + X\cdot T\cdot P_{fc} \tag{8}$$

which with (2) express the equilibrium

[7] The concept of a value-added product basic to such a supply function is explored by Chulsoon Khang.

[8] An equivalent procedure for analytical purposes would be to assume that all production is for *U.S.* consumption, i.e., that products of the type imported under *OAP* are consumed only in the United States.

price relationships of the model, and allow the removal of the $OAP$ (i.e., the comparative statics shift from the "with $OAP$" to the "without $OAP$" case) to be handled by taking the derivative of the model with respect to $X$ ($T$ held constant) and setting $\Delta X=1$.

The behavioral relations of the model are $Hd(P_{da})$ and $Hf(P_{fa})$, the domestic and foreign supplies of assembly; $Gd(P_{dc})$ and $Gf(P_{fc})$, the domestic and foreign supplies of components; and $K(P_j)$, the domestic demand for the finished good.

The total supplies of assembly and components, expressed by substituting for $P_{da}$ and $P_{dc}$ from (7) and (8) as functions of foreign prices are

$$(9) \quad Q_t = G_d[(1+(1-X)T)P_{fc}] + G_f[P_{fc}]$$

$$(10) \quad Q_t = H_d[(1+T)P_{fa} + XTP_{fc}] + H_f[P_{fa}]$$

Note that in (9) and (10) we do not distinguish between the total quantities of assembly and of components. This means that (9) and (10) incorporate the equilibrium condition that these two quantities are equal. Further $Q_t=K(P_j)$ is the equilibrium condition that quantity supplied of $J$ equals quantity demanded.

By definition, $P_j=P_{da}+P_{dc}$. When we substitute for $P_{da}$ and $P_{dc}$ from the equilibrium price relations (7) and (8), we have

$$(11) \quad P_j = (1+T)(P_{fa}+P_{fc})$$

When (11) is substituted into the demand function, we have

$$(12) \quad Q_t = K[(1+T)(P_{fa}+P_{fc})]$$

The initial system contains five behavioral relations (the two $H$, two $G$, and $K$), three definitions (of $P_{da}$, and of the total quantities supplied of assembly and of components), and four equilibrium conditions (equations (7) and (8), plus the implicit equality of the quantity demanded of $J$ with the total quantities supplied of assembly and of components). The variables in this system are $P_{da}$, $Q_{da}$, $P_{dc}$, $Q_{dc}$, $P_{fa}$, $Q_{fa}$, $P_{fc}$, $Q_{fc}$, $P_j$, the quantity demanded of $J$, and total quantities supplied of assembly and of components. In use this system is reduced to the three simultaneous equations (9), (10), and (12) in the variables $P_{fa}$, $P_{fc}$, and $Q_t$. Conceptually, after equilibrium values of these variables are solved for simultaneously, they can be substituted into (7), (8), (11) and the supply functions in order to determine equilibrium values of the other variables.

## II. Effects of the Offshore Assembly Provision

In this model the effects of the $OAP$ are the derivatives with respect to $X$ ($T$ held constant) of the system of equations. The resulting expressions will be evaluated for the case in which $X=0$, initially, but $T>0$. (An $OAP$ cannot be defined when $T=0$.)

After substituting (11) into (12), taking total derivatives of equations (9), (10), and (12), with $T$ held constant, and converting to elasticity terms, we have[9]

$$(13) \quad Q_t^* = \Phi_{tc}P_{fc}^* - \Phi_{dc}R_{dc}\cdot T/(1+T)\cdot dX$$

$$(14) \quad Q_t^* = \Phi_{ta}P_{fa}^* + \Phi_{da}R_{da}(T/(1+T)) \cdot(C/(1-C)dX)$$

$$(15) \quad Q_t^* = (1-C)\Gamma P_{fa}^* + C\Gamma P_{fc}^*$$

### A. *Effects on Prices and Outputs*

The solutions for $P_{fc}^*$ and $P_{fa}^*$ are shown in (16) and (17). If we replace the $dX$ by $\Delta X$ and let $\Delta X=1$, these expressions measure the effects of removing the existing $OAP$, and the assumptions that the $\Phi$ are positive and that $\Gamma$ is negative imply that

[9] From the total supply of components and of assembly functions (9) and (10) we can show that $\Phi_{ta}=R_{da}\cdot\Phi_{da}+R_{fa}\cdot\Phi_{fa}$, and $\Phi_{tc}=R_{dc}\cdot\Phi_{dc}+R_{fc}\cdot\Phi_{fc}$, where $R_{da}=Q_{da}/Q_t$, etc.

$$(16)\quad P^*_{fc} = \left[\frac{-\Gamma[C\cdot\Phi_{da}\cdot R_{da} + (1-C)\Phi_{dc}\cdot R_{dc}] + \Phi_{ta}\cdot\Phi_{dc}\cdot R_{dc}}{-\Gamma[C\cdot\Phi_{ta} + (1-C)\Phi_{tc}] + \Phi_{ta}\cdot\Phi_{tc}}\right]\left(\frac{T}{1+T}\right)dX$$

$$(17)\quad P^*_{fa} = \left[\frac{\Gamma[C\cdot\Phi_{da}\cdot R_{da} + (1-C)\Phi_{dc}\cdot R_{dc}] - \Phi_{tc}\cdot\Phi_{da}\cdot R_{da}}{-\Gamma[C\cdot\Phi_{ta} + (1-C)\Phi_{tc}] + \Phi_{tc}\cdot\Phi_{ta}}\right]\left(\frac{C}{1-C}\right)\left(\frac{T}{1+T}\right)dX$$

$P^*_{fc}>0$ and $P^*_{fa}<0$.

From equations (7), (8), and (11) we derive

$$(18)\quad P^*_{dc} = P^*_{fc} - \left(\frac{T}{1+T}\right)dX$$

$$(19)\quad P^*_{da} = P^*_{fa} + \left(\frac{T}{1+T}\right)\left(\frac{C}{1-C}\right)dX$$

Comparing the numerator and the denominator in both (16) and (17), we see that term by term the numerator of the expression in the square brackets is less in absolute value than the denominator. (For example, $\Phi_{da}\cdot R_{da}<\Phi_{ta}$, $\Phi_{dc}\cdot R_{dc}<\Phi_{tc}$, etc.) The expressions in square brackets in (16) and (17) are thus between 0 and 1 in absolute value, so (18) and (19) imply that $P^*_{dc}<0$, and $P^*_{da}>0$. The solutions for $Q^*_{da}$, $Q^*_{dc}$, $Q^*_{fa}$; $Q^*_{fc}$, have the same sign as the respective price changes, hence there remains only the question of the effect on $P_j$ and $Q_t$ of removing the $OAP$. The expression for the effect on $Q^*_t$ is

$$(20)\quad Q^*_t = \Gamma\cdot C\left[\frac{\Phi_{ta}\cdot\Phi_{dc}\cdot R_{dc} - \Phi_{tc}\cdot\Phi_{da}\cdot R_{da}}{den}\ \frac{T}{1+T}\right]dX$$

The denominator inside the square brackets is the same as in equations (16) and (17).

On closer examination of this expression we find that it is positive or negative as

$$\left(\frac{\Phi_{fa}}{\Phi_{da}}\right)\cdot\left(\frac{R_{fa}}{R_{fc}}\right)\cdot\left(\frac{\Phi_{dc}}{\Phi_{fc}}\right)\cdot\left(\frac{R_{dc}}{R_{da}}\right)$$

is less than or greater than 1.

Because $OAP$ imports contain domestic components and are assembled abroad, we know that $Q_{fc}<Q_{fa}$ and $Q_{da}<Q_{dc}$, i.e., $R_{fc}<R_{fa}$ and $R_{da}<R_{dc}$. This information about the relative positions of the domestic and foreign supply curves does not specify their relative elasticities. But because "assembly" is an activity with a relatively large input of unskilled labor, it seems reasonable to assume that expansion of assembly would require less bidding away of labor from higher paying activities if that expansion took place outside of the United States, i.e., to assume that $\Phi_{fa}>\Phi_{da}$. One is tempted to reason conversely that $\Phi_{fc}<\Phi_{dc}$. The foreign supply of components refers however not only to the labor-rich areas which might have comparative advantage in assembly, but also to industrial countries such as Japan and Germany, and the foreign supply curve is net of foreign demand, hence it is possible that $\Phi_{fc}>\Phi_{dc}$.

We cannot therefore conclude on a priori grounds that $Q^*_t$ is either positive or negative. But because $R_{fa}/R_{fc}$ and $R_{dc}/R_{da}$ are definitely greater than 1, and $\Phi_{fa}/\Phi_{da}$ is likely greater than 1, it seems likely that $Q^*_t/dX$ will be negative and $P^*_t/dX=(1/\Gamma)$. $(Q^*_j/dX)$ will be positive, i.e., that repeal of the $OAP$ will raise the price and reduce domestic consumption of goods imported under the provisions.

### B. *Effect on the Trade Balance*

The supply functions of the model all refer to production of the final good for sale in the U.S. This means that the net value of imports $M$ is equal to $P_{fc}Q_{fc}+P_{fa}Q_{fa}$. From this,

$$(21)\quad M^* = A(1+\Phi_{fa})P^*_{fa} + (1-A)(1+\Phi_{fc})P^*_{fc}$$

where $A$ is the value of assembly as a proportion of net imports. The first term on the right-hand side of (21) measures the effect of the $OAP$ on imports of $FA$; the second term its effect on imports of $FC$.

Because $P^*_{fa}<0$ and $P^*_{fc}>0$, it is clear that the change in the trade balance resulting from elimination of the $OAP$ will not necessarily be positive—the change being larger (in a positive sense) as the elasticities of supply of assembly in the two countries are smaller and as the elasticities of supply of components are larger.

## III. Welfare Effects

Following Harry Johnson's classic article, the net domestic welfare effect of a change in a trade policy variable is the sum of the change of consumers' surplus, the change of producers' surplus, and the transfers of wealth between the domestic and foreign economies associated with the collection of tariff revenue.

### A. *Change of Consumers' Surplus*

The change of consumers' surplus resulting from a policy change is always measured by the change of the area under the demand curve which is above the price paid by the buyer. Thus the gain of consumers' surplus is measured by

$$(-dP_j)Q_t + 1/2(-dP_j)(dQ_t)$$

or in relative terms,

$$(22)\quad \frac{Gain\ of\ CS}{P_jQ_t} = -P_j^*(1 + 1/2\Gamma\cdot P_j^*)$$

where $P_j$, $Q_t$ are the initial values of the variables and the $dP_j$, $dQ_t$ are the results of the change in the policy variables in question.

### B. *Change of Domestic Producers' Surplus*

As an $OAP$ and tariffs with and without $OAP$ will affect the level of $DA$ and of production of $DC$, there will be changes of producers' surplus in each sector. In each sector, the gain of producers' surplus will be equal to

$$(dP)Q + 1/2\,(dP)\cdot(dQ)$$

or, in relative terms

$$(23)\quad \frac{Gain\ of\ PS}{PQ} = P^*(1 + 1/2\,\Phi\cdot P^*)$$

where again $P$ and $Q$ represent initial values, and refer along with $\Phi$ to $DC$ in one case and to $DA$ in the other.

### C. *Transfers Associated with Tariff Revenue*

Figures 1 and 2 will be used to isolate the wealth transfers between the domestic and foreign economies which are associated with the $OAP$. In these figures, the first subscript distinguishes among domestic, foreign, and total (the sum of domestic and foreign) while the second distinguishes among the three policy situations—free trade ($o$), a tariff with $OAP$ ($p$), and a tariff without $OAP(t)$. To reduce the number of lines, the supplies of $DC$ and of $FC$ in Figure 1, and the supplies of $DA$ and $FA$ in Figure 2 have been assumed to be identical.

Because a tariff on $J$ without $OAP$ does not affect the relation between $P_{dc}$ and $P_{fc}$ (see equation (7) with $X=1$), such a tariff will not shift any of the curves in Figure 1. A tariff with an $OAP$ (equation (7) with $X=0$) shifts the supply of $FC$ and hence the total supply of components upwards.

The effects of removing the $OAP$ (with the tariff rate constant) are thus shown by moving from policy "$p$" to policy "$t$." With the $OAP$, the prices in Figure 1 of $DC$ and of $FC$ are $P_{dp}$, and the tariff revenue collected on imports of components is represented by the area $ADFC$. When the $OAP$ is eliminated, there is no longer any tariff revenue on components. The parts of former tariff revenue represented by rectangles $ADEB$ and $NEFM$ are transferred to other domestic uses (involving no change of welfare), while the part repre-

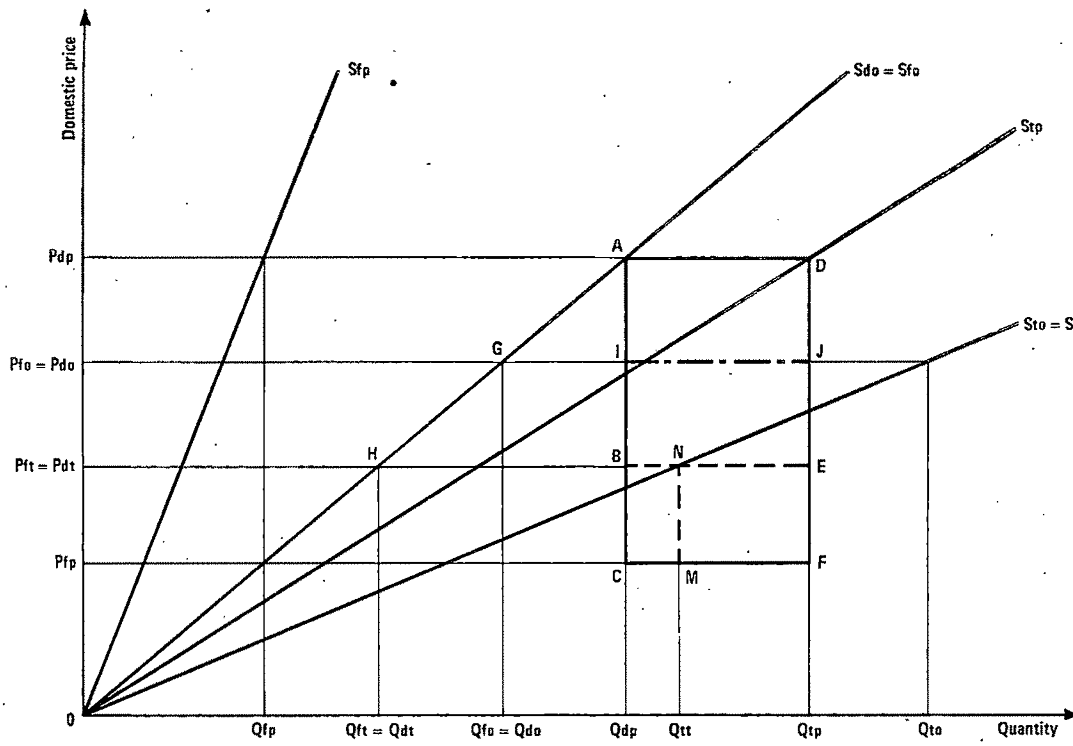



FIGURE 1. EFFECTS ON COMPONENTS

sented by *BNMC* is transferred from tariff revenue to payment for imports—a loss of domestic welfare.

The vertical dimension of the area *BNMC* is the change of the price of *FC* caused by eliminating the *OAP*, while the horizontal dimension is $Q_{tt}$ less $Q_{dp}$ which is equal to $Q_{ft}+Q_{dt}-Q_{dp}$. Changing the sign so that the change is defined as a gain rather than a loss, and converting to relative terms, we have

$$(24)\qquad \frac{Gain}{V_{fc}} = \frac{R_{dc}}{R_{fc}}\left(\frac{1}{1+T}\right)P^*_{fc}Q^*_{dc} - P^*_{fc}(1+Q^*_{fc})$$

In this expression $V_{fc}$ is the value in the $p$ instance, i.e., the instance for which we have observations.

Similarly, in Figure 2 the effects on *DA* of a tariff with and without an *OAP* have been sketched. As equation (8) shows, a tariff with an $OAP(X=0)$ shifts the supply of *FA* leftward, while a tariff without an $OAP(X=1)$ shifts it even further to the left. Tariff revenues in the $t$ and $p$ cases are represented by areas *ABCD* and *EFGH*. When the *OAP* is eliminated, area *IJKL* remains tariff revenue and hence does not involve a welfare change. Area *EILH* is transferred from tariff revenue to suppliers of *DA*, area *EIHL* is shifted from tariff revenue to suppliers of *DA*, area *ABJI* from domestic consumers to tariff revenue and area *JFGK* from tariff revenue to domestic consumers—involving in none of the three instances a change of the total wealth of the domestic

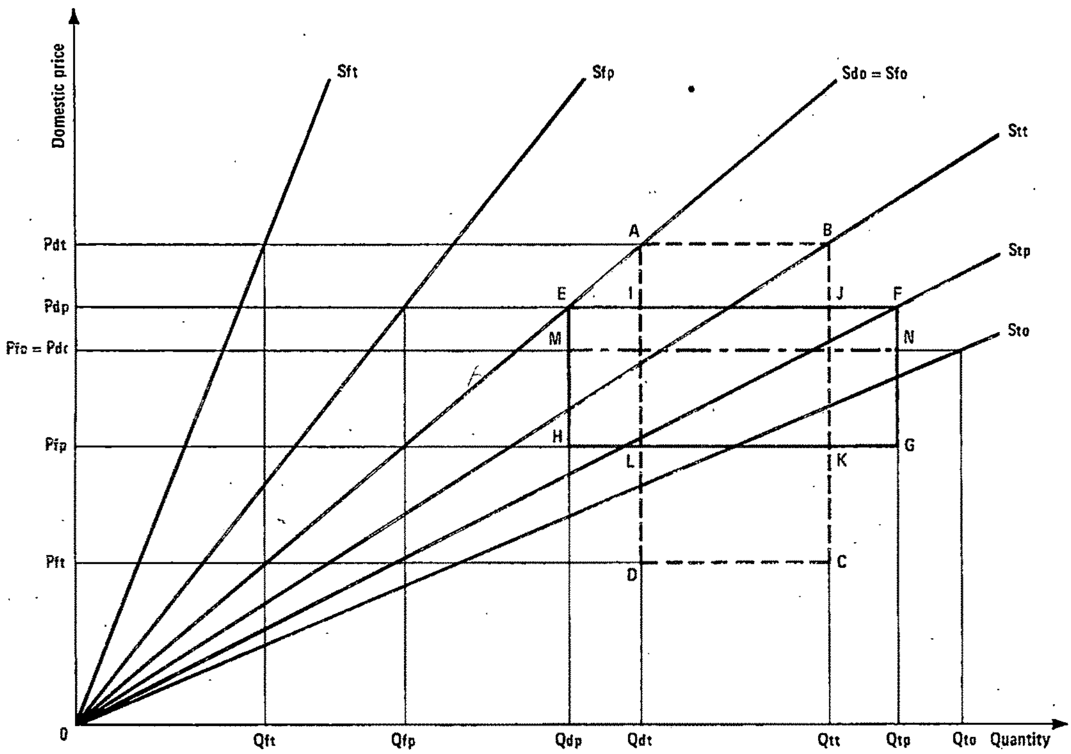



FIGURE 2. EFFECTS ON ASSEMBLY

economy.

Finally, *LKCD* is shifted from receipts of foreign assemblers in exchange for *FA* to tariff revenue—for which foreign assemblers receive no *quid pro quo.* Hence the area *LKCD* represents a welfare gain to the domestic economy. The horizontal and vertical dimensions of this area are $Q_{fa}$ in the *t* case, and the change of $P_{fa}$ caused by eliminating the *OAP*. Converted to relative terms, we have

$$\frac{Gain}{V_{fa}} = - P^*_{fa}(1 + Q^*_{fa}) \tag{25}$$

in which $V_{fa}$ is the value of *FA* in the *p* case.

## IV. Data Inputs and Parameters

### A. *Data*

The basic data used in the study were from the Census Bureau's report of the total and dutiable values of *OAP* imports. These data, plus data for "Directly Competitive Imports" (non-*OAP* imports in those 7-digit *TSUS* categories in which there were *OAP* imports)[10] provided the measures of imports of foreign assembly and of foreign components, and of *C*, the components' input-output coefficient.[11]

Components are identified in the data and therefore in the results only by the industry in which they are used, and not by the industry in which they are produced. It follows that labor-output coefficients could not be identified, and the

[10] I wish to thank the U.S. Tariff Commission and George N. Ecklund, Director of the Office of Economic Research at the Commission, for making these tabulations available.

[11] Tariff data were taken from a GATT tabulation, and import shares data (the parameters $R_{da}$ and $R_{fa}$) were based on data tabulated by the Trade Relations Council. Details of how the data were fitted together are available on request.

estimates of output displacement could not be translated into employment displacement estimates.

### B. *Parameters*

The selection of supply and demand elasticities began with estimates (from the literature) of the elasticity of import demand for finished goods ($\Gamma_m$). This elasticity was decomposed into consistent values for the elasticities of domestic demand ($\Gamma$) for and supply ($\Phi_s$) of finished goods, and the value of $\Phi_s$ was then decomposed into consistent values of $\Phi_{da}$ and $\Phi_{dc}$. Values for $\Phi_{fa}$ and $\Phi_{fc}$ were obtained by similar decomposition of the estimate of the foreign elasticity of export supply of finished goods used by Stephen Magee.[12]

The parameter values on which the results in Tables 1 and 2 are based are[13] $\Gamma=-0.75$, $\Phi_{da}=1$, $\Phi_{dc}=3$, $\Phi_{fa}=40$, and $\Phi_{fc}=10$. A sensitivity analysis indicated that the major conclusions reported below are quite robust with respect to assumed values for these parameters. The same values for these parameters were used for all industries, but the values for $T$, $C$, $R_{da}$, and $R_{dc}$ were specific to the industry.

## V. Results

The results shown in the tables are based on the assumption that $\Phi_{fa}$ and $\Phi_{fc}$ for textile products are restricted to zero by import quotas. Initial calculations using the same parameter values as for the other industries indicated that *OAP* removal would cause *U.S.* imports of foreign assembly of textile products to become negative—i.e., would cause the United States to become a reexporter of textile components produced abroad but assembled in the United States. This result seems to be inconsistent with the generally accepted pattern of *U.S.* comparative advantage and disadvantage, and might be the result of basing the estimates on a situation in which imports of assembly are less than they would be if all variables were at long-run equilibrium values, as the model assumes.

Comparison with the list of textile products subject to import quota[14] indicated an overlap with virtually every category of *OAP* imports, hence the assumption that $\Phi_{fa}$ and $\Phi_{fc}$ are zero for the textile industry. Because this assumption is somewhat tenuous, separate totals with and without textile products, which account for less than 3 percent of *OAP* imports, are given.

Because the ratio of duty-free value (i.e., domestic components) to total value, as given in the second column of Table 1, is very low for motor vehicles and household sewing machines, the tariff saving from the *OAP* is very low and hence the effects of the *OAP* on these industries is slight. In terms of gross value of trade involved, motor vehicles account for almost 40 percent of *OAP* imports. But in terms of the effects of the *OAP*, this industry is an almost insignificant part.

With motor vehicles and household sewing machines (on which the effects of the *OAP* are very slight) excluded, about half the value of imports affected by the *OAP* is assembly and half is foreign components. Thus the effects of the *OAP* on

[12] The reason for using indirect estimates is that more direct ones are not available. The relations involved are

$$\Phi_m = 1/M(\Gamma - \Phi_s) + \Phi_s$$

and

$$1/\Phi_s = ((1 - C)/\Phi_{da}) + C/\Phi_{dc}$$

The first of these is given by Magee and the derivation of the second is available from the author on request. The second expression, with the subscripts appropriately modified, relates $\Phi_{fa}$ and $\Phi_{fc}$ to the foreign elasticity of export supply of the finished good.

[13] This value for $\Gamma$ was used by Magee. The values for $\Phi_{fa}$ and $\Phi_{fc}$ are consistent (by the expression in fn. 12) with Magee's value of 1.5 for $\Phi_s$. The values of $\Phi_{fa}$ and $\Phi_{fc}$ imply a foreign export supply elasticity of the finished good equal to 16, which is slightly higher than the value of 14.4 used by Magee.

[14] See U.S. Department of Commerce.

TABLE 1—VALUES OF IMPORTS AFFECTED BY THE OFFSHORE ASSEMBLY PROVISION, 1972, AND ESTIMATED EFFECTS ON THE TRADE BALANCE OF THE REMOVAL OF THE OFFSHORE ASSEMBLY PROVISION

(Values in millions of dollars)

| | *OAP* Imports | | | | Effects on imports of removing *OAP* | | |
|---|---|---|---|---|---|---|---|
| Industry | Total value | Duty free value as percent of total value | Tariff rate[a] | Directly competitive imports | Assembly | Components | Total |
| Textile products | 87 | 52 | 26 | 1202 | −60 | 82 | 23 |
| Lumber and paper products | 9 | 63 | 7 | 217 | −61 | 52 | −9 |
| Household sewing machines | 31 | 1 | 6 | 79 | −0.0 | 0.1 | 0.1 |
| Office machinery | 251 | 50 | 5 | 302 | −98 | 62 | −36 |
| Electronic products | 607 | 37 | 7 | 1093 | −122 | 157 | 35 |
| Motor vehicles | 1310 | 2 | 4 | 3488 | −24 | 19 | −5 |
| Aircraft | 109 | 34 | 5 | 365 | −86 | 50 | −35 |
| Other machinery | 557 | 28 | 6 | 5355 | −688 | 616 | −72 |
| Scientific instruments | 17 | 43 | 9 | 156 | −34 | 36 | 2 |
| Sporting goods | 45 | 38 | 14 | 207 | −38 | 52 | 14 |
| Miscellaneous manufactures | 69 | 43 | 13 | 2638 | −844 | 865 | 21 |
| Metal products (item 806.30) | 318 | 59 | 7 | 2247 | −715 | 617 | −98 |
| Total | 3409 | 25 | | 17,350 | −2768 | 2608 | −160 |
| Total excluding textiles | 3322 | 25 | | 16,148 | −2709 | 2526 | −183 |

*Note:* Parameter values are $\Gamma=-.75$, $\Phi_{da}=1$, $\Phi_{dc}=3$, $\Phi_{fa}=40$, $\Phi_{fc}=10$, for all except textile products. For these products, $\Phi_{fa}$ and $\Phi_{fc}$ were assumed restricted to zero by quotas.

[a] Shown in percent.

assembly and on components have a tendency to cancel each other.

### A. *Balance of Trade*

It appears that the balance-of-trade effect of the *OAP* is rather slight. With the parameter values used in Table 1, I estimate that repeal of the *OAP* would have reduced net-imports by less than $200 million in 1972. The estimated increase of imports from repeal of the *OAP* is less than 2 percent of total affected imports (dutiable value of *OAP* imports plus directly competitive imports)—this small net effect resulting from the offsetting changes of imports of assembly and components. By contrast, Magee estimated that removal of the tariff on imports directly competive with *U.S.* production would cause these imports to increase by 65 percent.

The conclusion that repealing the *OAP* would not significantly improve the trade balance was confirmed by the results of the sensitivity analysis. When other "reasonable" parameter values were used, it was often estimated that removal of the *OAP* would increase net imports.

### B. *Domestic Effects*

Estimates of the effects of removal of the *OAP* on domestic consumption and output, and on domestic economic welfare are given in Tables 2a and 2b. For perspective I have listed the value of domestic shipments and the implicit value of domestic output competitive with *OAP* imports,[15] calculated by dividing $V_{fa}$ and $V_{fc}$

[15] Because the $R_{fa}$ and $R_{fc}$ values used were matched as

TABLE 2A—VALUE OF DOMESTIC PRODUCTION DIRECTLY COMPETITIVE WITH *OAP* IMPORTS

(Values in millions of dollars)

| Industry | Value of domestic shipments 1971 (a) | Implicit value in 1972 of domestic output competitive with products imported under *OAP* Total (b) | Assembly (c) | Components (d) |
|---|---|---|---|---|
| Textile products | 27,973 | 4,707 | 2,037 | 2,670 |
| Lumber and paper products | 22,178 | 2,067 | 1,018 | 1,049 |
| Household sewing machines | 143 | 72 | 71 | 1 |
| Office machinery | 7,021 | 3,929 | 1,974 | 1,955 |
| Electronic products | 24,944 | 4,488 | 2,767 | 1,721 |
| Motor vehicles | 60,845 | 28,830 | 28,239 | 591 |
| Aircraft | 15,203 | 5,768 | 3,863 | 1,905 |
| Other machinery | 72,287 | 42,373 | 31,122 | 11,251 |
| Scientific and precision instruments | 12,311 | 967 | 576 | 391 |
| Sporting goods, toys, games | 2,786 | 829 | 534 | 295 |
| Miscellaneous manufactures | 21,470 | 19,363 | 11,729 | 7,634 |
| Metal products (item 806.30) | 42,026 | 20,970 | 8,887 | 12,083 |
| Total | 309,214 | 134,363 | 92,817 | 41,546 |
| Total, excluding textiles | 281,241 | 129,656 | 90,780 | 38,876 |

TABLE 2B—ESTIMATES OF THE EFFECTS ON THE DOMESTIC ECONOMY OF REMOVAL OF THE OFFSHORE ASSEMBLY PROVISION

| Industry | Quantity of $J$ consumed[a] (e) | Assembly Quantity[a] (f) | Assembly Value[b] (g) | Components Quantity[a] (h) | Components Value[b] (i) | Total Quantity[a] (j) | Total Value[b] (k) | Net Welfare Gain[b] (l) |
|---|---|---|---|---|---|---|---|---|
| Textile products | 0 | 0 | 0 | 0 | 0 | 0 | 0 | −23.3 |
| Lumber and paper products | −1.1 | 5.0 | 102 | −6.0 | −84 | −0.6 | 18 | −7.2 |
| Household sewing machines | −0.01 | 0.05 | 0.1 | −13.3 | −0.1 | −0.1 | −0.1 | −0.03 |
| Office machinery | −1.1 | 3.9 | 155 | −4.1 | −108 | −0.1 | 4.7 | −12.9 |
| Electronic products | −0.8 | 3.5 | 192 | −9.9 | −227 | −1.6 | −35 | −36.6 |
| Motor vehicles | −0.03 | 0.06 | 31 | −3.0 | −24 | −0.00 | 8 | −3.3 |
| Aircraft | −0.06 | 1.6 | 127 | −3.1 | −80 | 0.05 | 47 | −7.8 |
| Other machinery | −0.5 | 1.7 | 1,042 | −5.9 | −883 | −0.2 | 209 | −95.4 |
| Scientific and precision instruments | −1.1 | 4.9 | 57 | −10.6 | −55 | −1.4 | 2 | −5.6 |
| Sporting goods, toys, games | −1.3 | 6.2 | 66 | −19.9 | −78 | −3.1 | −12 | −8.7 |
| Miscellaneous manufactures | −1.5 | 6.2 | 1,460 | −13.4 | −1359 | −1.5 | 100 | −128.7 |
| Metal products (item 806.30) | −1.3 | 6.9 | 1,231 | −6.4 | −1032 | −0.8 | 199 | −88.6 |
| Total | −0.7 | 2.4 | 4,464 | −7.1 | −3931 | −0.5 | 533 | −418.3 |
| Total, excluding textiles | −1.0 | 2.5 | 4,464 | −7.6 | −3931 | −0.5 | 533 | −395.0 |

[a] Shown in percent.
[b] Values in millions of dollars.

closely as possible with the product composition of imports affected by the *OAP*, they were usually larger than such figures for the entire industry. For this reason, and because imports affected by the *OAP* do not include all imports, the implicit values of affected domestic production are smaller than the values of domestic shipments of each industry.

by $R_{ja}$ and $R_{jc}$, respectively. The estimate for each industry of the effect on total domestic real output (col. (j)) is the average of the changes in real output of assembly and components, weighted by the input-

output coefficients (measured at domestic prices) for assembly and components. The other estimates, of course, were made by substitution of values into the appropriate expressions from Sections III and IV. For the estimates of changes of real output (cols. (f), (h), (j)) the "Total" lines are averages of the industry estimates weighted by the corresponding values from columns (b), (c), or (d).

The estimates indicate rather strongly that if the *OAP* were repealed the output of components would decline sufficiently to more than offset the increase of domestic assembly, so that the net impact would be a slight reduction of real domestic economic activity. Among product groups, the estimated net impact on domestic activity of removal of the *OAP* is positive only for aircraft, and in this case the figure is very small, only 0.05 percent.[16]

The estimated percentage effects on domestic output are relative to the implicit value of domestic output in column (b) rather than to value of total domestic shipments, as given in column (a). The 0.5 percent reduction thus is 0.5 percent of \$134 billion or about \$672 million (at 1972 prices). By comparison, the total value of manufacturers' shipments in 1971 was about \$671 billion.

Repeal of the *OAP* would reduce consumption of products affected by it by about 1 percent, and, as demand is assumed inelastic, would cause consumer expenditure on these products to increase by about \$200 million.

Finally, the welfare calculation in the last column of Table 2b indicates that the *OAP* generated in 1972 a net welfare benefit of about \$400 million. The net loss which would result from its repeal is made up of a \$1,424 million loss of consumers' surplus and a \$346 million transfer loss associated with tariff revenues, partly offset by a gain of producers' surplus of \$1,349 million. The sensitivity analysis again shows that the choice of parameters was not critical.

## VI. Offshore Assembly Provisions as Development Incentives

One is on solid ground to conclude that the United States benefits in a welfare sense from the existence of the *OAP*. I have pointed out elsewhere that in 1972 the value of *U.S.* imports from less developed countries (*LDC*) under the *OAP* was as large as exports of these countries to the European Economic Community (*EEC*) under the *EEC*'s system of tariff preferences, and I have estimated that the *U.S. OAP* generated in 1972 net foreign exchange earnings for *LDC* equal to those generated by the preference system of the *EEC*. And, while *U.S. OAP* imports from the *LDC* have been increasing very rapidly, it has been pointed out frequently that the potential for growth of *LDC* preferential exports to the *EEC* is severely restricted by the ceilings and quotas built into the scheme.[17] Hence it is probable that the *OAP* scheme of the United States at present is a more important contributor to the foreign exchange earnings of the *LDC* than is the *EEC*'s preference scheme, and that the difference is increasing rapidly year by year.

Why then do we observe these tight restrictions placed on preferential imports (indeed, the United States has only recently implemented a limited preference scheme), while *OAP* imports into the United States continue to increase rapidly? To the exporting country, preferences or *OAP* treatment are similar incentives.[18] Tariff preferences allow the *LDC* to compete more effectively with domestic producers in the industrial countries on those

[16] The conclusion that *OAP* removal will reduce domestic activity is reinforced by the sensitivity analysis.

[17] See Richard Cooper and Tracy Murray.

[18] As the *EEC* preference scheme allows generally for 50 percent preferences, and the dutiable value of *U.S. OAP* imports is about 50 percent of total value, the tariff reductions allowed by the two schemes are about equal.

products in which the *LDC* have comparative advantage—provided that these products are not excluded from the preference scheme and that the *LDC* are not required to use components produced domestically at high cost in order to satisfy the "country of origin" requirements of the preference scheme.

As for an *OAP*, it allows the labor rich *LDC* to do the assembly jobs in which they have comparative advantage without subjecting them to the higher effective rates of protection they would face if forced to bear the duty on the value of components which they do not produce.

Of course, this "out" provided by the *OAP* has a catch in it. The *LDC* must use *U.S.* components in order to qualify for the tariff reduction. But the estimates of the impact of the *OAP* on *LDC* exports referred to above indicate that this out with a catch is better than no out at all.

The other side of the *OAP*—its negative impact on production of foreign components—is probably not extremely burdensome to *LDC*. The displacement of foreign components by *U.S.* components in the *LDC* assembling country will often be the replacement by *U.S.* components of Japanese or German components rather than the displacement of components produced by *LDC*. Hence the shifting of the assembly process to the *LDC* will often be a net as well as a gross gain to these countries.

The fact that the *OAP* shifts demand toward domestic components generates domestic political support in the producing sector, which means that an *OAP* is more likely to attract sufficient political push for enactment and expansion than a tariff preference scheme, which has only widely diffused consumer interests and good will to back it.[19]

This suggests that efforts to liberalize *OAP* treatment of articles assembled in *LDC* could produce considerably greater benefits to the *LDC* than have resulted from attempts to negotiate tariff preferences.

REFERENCES

**R. N. Cooper**, "The European Economic Community's System of Generalized Tariff Preferences: A Critique," *J. Develop. Stud.*, July 1972, *8*, 379–94.

**J. M. Finger**, "Tariff Provisions for Offshore Assembly and the Exports of Developing Countries," *Econ. J.*, June 1975, *85*, 365–71.

**H. G. Johnson**, "The Cost of Protection and the Scientific Tariff," *J. Polit. Econ.*, Oct. 1960, *48*, 327–45.

**C. Khang**, "Factor Substitution in the Theory of Effective Protection," *J. Int. Econ.*, Aug. 1973, *3*, 227–46.

**S. P. Magee**, "The Welfare Effects of Restrictions on U.S. Trade," *Brookings Papers* Washington 1972, *3*, 645–707.

**T. Murray**, "Preferential Tariffs for the LDCs," *Southern Econ. J.*, July 1973, *40*, 35–46.

**Economist Intelligence Unit**, "Electronic Components," *Multinational Business*, No. 2, 1973, 23–31.

**GATT**, *Basic Documentation for Tariff Study, Summary Table No. 2—Tariff and Trade Profiles by Product Categories,* Geneva 1970.

**Trade Relations Council of the United States, Inc.**, *Employment, Output, and Foreign Trade of U.S. Manufacturing Industries, 1958–71,* 5th ed., 3 vols., Geneva 1973.

**U.N. Conference on Trade and Development**, *Second General Report on the Implementa-*

[19] Additional support for an *OAP* comes from domestic firms who shift assembly operations to offshore affiliates, but these firms would also benefit from tariff reductions and hence would be expected to also support preferences.

Data are not available which would separate *OAP* imports among those assembled in offshore affiliates of *U.S.* firms, foreign assemblers who work under contract with *U.S.* importers, and foreign firms which buy *U.S.* components on their own account for assembly into product which will be marketed in the United States. The Tariff Commission's report, p. 94, does however state that two-fifths of *OAP* imports of apparel came from foreign establishments in which no financial interest was held by *U.S.* firms.

*tion of the Generalized System of Preferences,* Doc. No. TD/B/C.5/22, Apr. 17, 1974, Geneva.

U.S. Bureau of the Census, *U.S. Imports of Merchandise for Consumption and General Imports of Merchandise,* rept. no. IA 245-A, Washington 1973.

U.S. Department of Commerce, *Long Term International Cotton Textile Arrangement Categories,* Washington 1968.

U.S. Tariff Commission, *Economic Factors Affecting the Use of Items 807.00 and 806.30 of the Tariff Schedules of the United States,* TC publication 339, Washington 1970.

# Urban Renewal Policy and Economic Efficiency

*By* LAWRENCE D. SCHALL*

It has been argued that slums can arise because property owners seek to maximize their own profits and ignore the impact of their actions (for example, upkeep investment) on the values of surrounding properties; this disregard of "technological" externalities, or spillovers, by owners can produce a slum equilibrium at which maintenance outlays and *total* neighborhood profits are less than would prevail were such externalities internalized (taken into account).[1] Contrary to this view, the arguments of Ronald Coase imply that, with no transaction costs, owners will agree to upgrade their structures to a level maximizing total (joint) profits. With transaction costs, however, it is clear that this may not occur. A common prescription is renewal through a government initiated rehabilitation or redevelopment project.

Existing treatments of externalities and renewal center almost entirely around the competitive market's failure and the state's ability through renewal to internalize the spillover effects. These discussions have ignored three important issues, each of which will be explored in the present paper:

(i) The role of externality relationships in the neighborhood in determining the possible renewal states that can be achieved and maintained by alternative renewal approaches.

(ii) The conditions under which a renewed neighborhood will not revert to a slum even though postrenewal ownership and decision making remain fragmented. If prerenewal failure to internalize externalities produced the slum, the same behavior may imply redeterioration after renewal. It is shown below that the likelihood of postrenewal decay depends on the externality relationships and on the renewal strategy; the externalities will determine whether permanent controls will be needed to sustain renewal under fragmented ownership.

(iii) The method for evaluating both the relative economic efficiency of alternative renewal goals and the techniques for achieving those goals.

In Section I, it is shown that several equilibrium solutions other than the initial slum equilibrium exist for a neighborhood. Each solution is a potential renewal goal. The equilibria that are attainable depend on the nature of the land use spillover effects within the neighborhood. In Section II, and the related appendices, it is shown that the major conclusions of Section I hold under more general assumptions. In Section III, the results are used to develop a general method for evaluating renewal policies. The analysis determines those renewal approaches or strategies that are *effective* in permanently attaining each alternative renewal goal; from the effective approaches, a subset of *efficient*, i.e., cost minimizing, renewal strategies is identified. The renewal approach that is used

* Graduate School of Business, University of Washington. I wish to thank an unknown referee for valuable comments.

[1] See, for example, Otto Davis and Andrew Whinston (1961), Hugh Nourse, and Jerome Rothenberg; see Martin Bailey and the author (1971a) on land use choices and technological externalities. These externalities involve a change in a plot's production possibilities due to an altered neighborhood environment and not due to input price changes ("pecuniary" externalities); on technological externalities, see Buchanan and Stubblebine, E. J. Mishan, and the author (1971c).

should be efficient and the preferred renewal goal (which may be no renewal at all) is that which generates the greatest net benefits given the use of an efficient strategy in achieving that goal. An implication of the analysis is that permanent renewal can be efficiently achieved with fragmented ownership and with no post-renewal controls (private or government) if more than one competitive equilibrium exists for the neighborhood; if no higher competitive equilibrium exists, renewal with fragmented ownership will be inefficient unless long-run controls are imposed.

## I. Neighborhood Spillover Relationships

Four levels of improvement for a neighborhood will be compared: a slum competitive equilibrium ($c$), a higher competitive equilibrium ($c'$), a "cooperative" equilibrium ($o$), and a "state solution" ($s$). A competitive equilibrium involves fragmented ownership of the neighborhood with each owner seeking to maximize his own returns, disregarding the impact of his actions on surrounding property values. Assuming the neighborhood is initially at slum equilibrium $c$, it is shown that a higher competitive equilibrium $c'$ may also exist for the neighborhood. At the cooperative equilibrium $o$, all externalities are taken into account and net returns to property owners are maximized over the entire neighborhood, for example, as would occur with single ownership of the neighborhood or with permanent joint planning by property owners under fragmented ownership. A state solution is an objective that takes into account factors in addition to the net returns of neighborhood property owners (for example, government revenues and income distribution). The government would have to compel owners to maintain a state solution.

It is assumed in this section that all plots in the renewal area are identical, that there is a single land use variable "quality" designated $Q$, and that the neighborhood is small enough so that changes in land use do not affect the prices of land use services in the general metropolitan area. In Section II (and related appendices) it is shown that essentially the same conclusions hold with these assumptions relaxed.

### A. *Competitive Equilibrium*

Assume the competitive equilibrium of a neighborhood with $n$ separate plots. Define $Q_i$ as a quality measure of the structure and grounds of plot $i$. Any level of $Q_i$ implies a particular set of physical characteristics for $i$ with a higher $Q_i$ requiring a greater dollar investment. Define function $I_i(Q_i)$ as the present value of the investment in property $i$ necessary to *achieve and maintain* property $i$ at any quality level $Q_i$ *given the initial competitive equilibrium*, $c$. The net present value of returns from plot $i$ equals

$$(1) \qquad V_i = R_i(Q_i, \overline{Q}_i) - I_i(Q_i)$$

where $R_i(Q_i, \overline{Q}_i)$ is the present value of the revenues (after tax) from plot $i$ and $\overline{Q}_i$ equals the average $Q$ on all other parcels, that is

$$(2) \qquad \overline{Q}_i = [1/(n-1)] \sum_{j \neq i} Q_j$$

$R_i$ therefore depends on the average quality level of other neighborhood properties. Each owner $i$ maximizes $V_i$ with respect to $Q_i$ for any given $\overline{Q}_i$. Setting $\partial V_i/\partial Q_i = 0$ in (1):

$$(3) \qquad \partial R_i/\partial Q_i = \partial I_i/\partial Q_i$$

Equation (3) states that the incremental return to $i$ from $Q_i$ equals the marginal cost of $Q_i$. Schedule $Q_i(c)$ (identical for all $i$ with identical plots) in Figure 1 indicates the level of $Q_i$ satisfying (3) for each level of $\overline{Q}_i$. At $c$, which is on the 45° line, $Q_i = Q_j = Q^c$, all $i, j$ (for proof see Appendix 1).

Slum equilibrium $c$ is stable. Thus, for

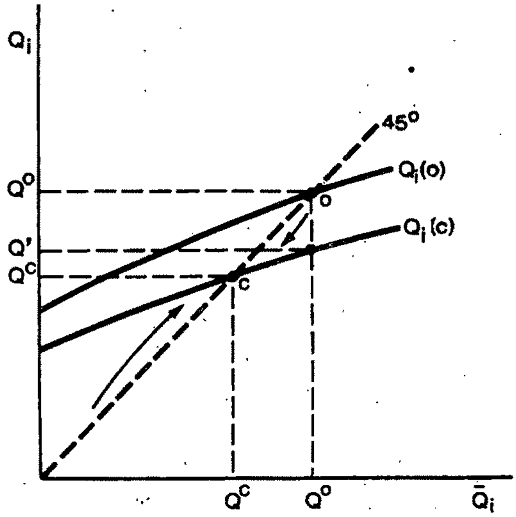



FIGURE 1

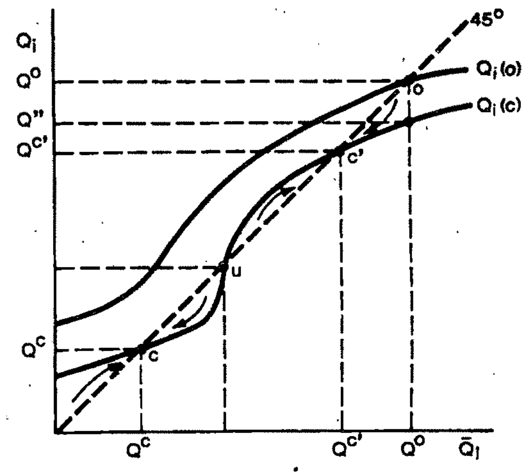



FIGURE 2

example, if $\bar{Q}_i$ were uplifted to $Q^o$, each owner $i$ would maximize his *own* profits by setting $Q_i = Q' < Q^o$. The result would be a deterioration of the neighborhood toward $Q'$. As $\bar{Q}_i$ declines, each owner would seek a $Q_i$ below $Q'$. Eventually the neighborhood would deteriorate to $c$.[2] This result holds for all $\bar{Q}_i > Q^c$ and the arrows in the figure indicate the direction of neighborhood quality change for each existing level of $\bar{Q}_i$. *Permanently uplifting the neighborhood above $Q^c$ must therefore provide for some means to compel competitive owners to maintain that higher $Q_i$.*[3]

Now assume the multiple competitive equilibria case of Figure 2. Equation (3) holds at locally stable competitive equilibria $c$ and $c'$, where $c$ still represents the slum equilibrium. Thus, if $\bar{Q}_i = Q^o$ in Figure 2, owners maximize their own profits by setting $Q_i = Q'' < Q^o$. Over time, the neighborhood declines to $c'$. Point $u$ in Figure 2 is an unstable competitive equilibrium.[4]

To clarify the multiple competitive equilibria concept, assume that all owners simultaneously upgrade their properties from slum level $Q^c$ to $Q^{c'}$. With the neighborhood renewed (rise in $\bar{Q}_i$), the marginal *private* product of investment by each owner has increased.[5] All owners conse-

[2] Only to simplify the discussion, the postrenewal slum is assumed to be the same as prerenewal slum $c$; in actuality, the two slums would differ.

[3] For $\bar{Q}_i < Q_i^c$, Figure 1 implies that owners will improve their properties above the existing $\bar{Q}_i$ until $Q_i^c$ is reached by all. This assumption is not necessary for the points to be made here; schedule $Q_i(c)$ might even lie along the 45° line for $\bar{Q}_i \leq Q_i^c$ in which case all $Q_i \leq Q_i^c$ would be equilibrium levels.

[4] A competitive equilibrium is stable if $0 \leq dQ_i(c)/d\bar{Q}_i < 1$, "system stable" (stable for all $\bar{Q}_i$) if $dQ_i(c)/d\bar{Q}_i = 1$, and unstable if $dQ_i(c)/d\bar{Q}_i > 1$; $c$ and $c'$ are locally stable and $u$ is unstable. To insure stability, it is assumed that all property owners maximize profits under the assumption that their neighbors will maintain their existing levels of $Q$; see Davis and Whinston (1962). It is assumed that $\partial^2 V_i/\partial Q_i^2 < 0$ at $c$ and $c'$, and, at all points, $\partial^2 V_i/\partial\bar{Q}_i\partial Q_i \geq 0$ and $\partial^2 V/\partial\bar{Q}_i\partial Q_i \geq 0$, where $V = \Sigma V_i$. Note that the multiple competitive equilibria case here differs from the Marshallian multiple equilibria for an industry with multiple intersections of a downward sloping demand and downward sloping supply. Prices here can be given even with multiple competitive equilibria, i.e., the area being renewed can be small enough not to alter the price structure for land use services.

[5] Differentiating competitive equilibrium condition $\partial V_i/\partial Q_i = 0$ (used for (3)), $dQ_i(c)/d\bar{Q}_i = -(\partial^2 V_i/\partial\bar{Q}_i\partial Q_i)/(\partial^2 V_i/\partial Q_i^2)$ on schedule $Q_i(c)$. For multiple competitive equilibria, $dQ_i(c)/d\bar{Q}_i > 1$ at some point between $c$ and $c'$;

quently find it profitable to maintain their structures at the new quality level ($Q^{c'}$) as long as their neighbors do likewise. With all owners behaving in this way, $c'$ is a self-sustaining locally stable competitive equilibrium. At $c'$, each owner is maximizing his own profits and is acting competitively, i.e., is *not* taking into account the externalities he is imposing on others. Also, although all $V_i$ are higher at $c'$ than at $c$ (see Section II), the value of the neighborhood is even higher at cooperative equilibrium $o$ ($o$ is discussed below) than at $c'$; at $o$, all marginal externalities are taken into account so as to maximize $V$ ($V=\Sigma V_i$). Nevertheless, as noted in Section III, $c'$ may be preferred to $o$ as a renewal goal because of lower upkeep enforcement costs.

Since all $V_i$ are higher at $c'$ than at $c$, an agreement among property owners (encouraged by the state) could be effected to raise the neighborhood from $c$ to $c'$. Or temporary state incentives (for example, an investment subsidy) could be instituted to produce $c'$. Only *temporary* measures are needed since $c'$ is a stable competitive equilibrium and once attained will be voluntarily sustained by property owners. Without state encouragement, joint property owner action may not arise, even though all would benefit, because of the transaction costs involved. No property owner may himself benefit sufficiently from renewal to initiate cooperative action.

A second competitive equilibrium above slum equilibrium $c$ is not implausible. A shift from $c$ to $c'$ is a transformation to a better neighborhood such as that which might exist (and be in competitive equilibrium) elsewhere in the metropolitan area. Just as most stable well-maintained neighborhoods are sustained without the imposition of upkeep restrictions, it is possible that the same situation is achievable through renewal in a currently deteriorated area. Indeed, there may be several higher competitive equilibria. Whether a $c'$ exists will depend upon such factors as the size of the renewal site, land uses surrounding the neighborhood, and the supply and demand for land use services throughout the metropolitan region.

### B. *Cooperative Equilibrium*

If owner $i$ were to take into account the effect of his investment on the value of the entire neighborhood, $V=\sum_{i=1}^{n} V_i$, then setting $\partial V/\partial Q_i=0$ to maximize $V$ with respect to $Q_i$ for any given level of $\overline{Q}_i$ implies that

$$(4)\qquad \partial R_i/\partial Q_i + \sum_{j\neq i} (\partial V_j/\partial Q_i) = \partial I_i/\partial Q_i$$

where, using (2),[6]

$$\begin{aligned}\partial V_j/\partial Q_i &= (\partial V_j/\partial \overline{Q}_j)(\partial \overline{Q}_j/\partial Q_i)\\ &= [1/(n-1)](\partial V_j/\partial \overline{Q}_j)\end{aligned}$$

Equation (4) states that the incremental return to the *entire neighborhood* from $Q_i$ equals the marginal cost of $Q_i$. Schedule $Q_i(o)$ is above $Q_i(c)$ in Figure 1 or Figure 2, is identical for all $i$, and indicates the level of $Q_i$ satisfying (4) for each level of $\overline{Q}_i$.[7] At cooperative equilibrium $o$ for the neighborhood, $V$ is maximized with respect to all $Q_i$ and (4) holds simultaneously for all

i.e., $\partial^2 V_i/\partial \overline{Q}_i \partial Q_i > -\partial^2 V_i/\partial Q_i^2$ where $\partial^2 V_i/\partial Q_i^2<0$ (by the strict concavity of $V_i$; see fn. 7); thus, the marginal profitability to $i$ of investment in $Q_i$, $\partial V_i/\partial Q_i$, must rise with $\overline{Q}_i$ and by more than the fall in $\partial V_i/\partial Q_i$ due to the increase in $Q_i$. This must hold for $i$ to be willing to increase $Q_i$ from $Q_i^c$ to $Q_i^{c'}$ and to maintain $Q_i^{c'}$.

[6] In this paper, it is assumed that $V=\Sigma V_i$, i.e., with uncertainty total value $V$ is independent of portfolio diversification effects from combining plot incomes; this is relevant with differing plots in Sections II to IV. By the principles of "risk independence" (Stewart Myers) and "value additivity" (the author, 1972a) $V=\Sigma V_i$ if capital markets (but not necessarily the neighborhood property market) are competitive and transaction costless. $V=\Sigma V_i$ may also hold with imperfections (see the author, 1971b).

[7] It is assumed in this section that an improved neighborhood environment raises a property's value, i.e., $\partial V_j/\partial \overline{Q}_j>0$; $V$ and $V_i$ are strictly concave functions of $Q_i$. Therefore, in Figures 1 and 2, schedule $Q_i(o)$ is above $Q_i(c)$ at all points and $Q^o>Q^c$ in Figure 1 and $Q^o>Q^{c'}$ in Figure 2.

plots. At $o$, $Q_i = Q_j = Q^o$, all $i$, $j$ (see Appendix 1).

Since total private (after tax) neighborhood returns are maximized at $o$, equilibrium $o$ will arise with permanent joint planning by property owners or with sole ownership of the neighborhood. Transaction costs of joint action or of effecting single ownership may prevent these outcomes, however. With retention of fragmented ownership, equilibrium $o$ will be permanently maintained only if *permanent* (and enforced) cooperative arrangements between property owners are effected or if the state imposes *permanent* controls to insure $o$. It follows from the earlier discussion that even if all owners uplift their properties to $Q^o$, unless compelled to remain there, each owner will permit his property to deteriorate to $Q'$ in Figure 1 (or to $Q''$ in Figure 2), resulting in an eventual regression to $c$ (or to $c'$ in Figure 2). Thus, with fragmented ownership, temporary owner cooperation or temporary state measures will not permanently maintain $o$.

### C. *State Solutions*

Any renewal objective for the neighborhood other than a competitive or cooperative equilibrium is referred to here as a "state solution" ($s$). A state solution may be preferred if it produces benefits not reflected in the returns to neighborhood property owners ($V$), for example, an improved income distribution, greater government revenues, or greater aesthetic returns to the city's inhabitants. A fundamental distinction between a state solution and the competitive and cooperative equilibria is that, in contrast with the latter outcomes, a state solution can be sustained over time only with *permanent government* intervention. This follows from the above discussion since, with fragmented ownership, $c$, $c'$, or $o$ will arise depending upon the degree of cooperation among owners; with sole ownership, $o$ arises. Therefore, the state must compel owners, for example, by permanent land use controls or fiscal measures (taxes or subsidies), to achieve and maintain any other position.[8]

## II. Generalization of the Model

Section I assumed identical plots, a single land use variable ($Q$), and no change due to renewal in the metropolitan area's price structure for land use services. Relaxing these assumptions does not nullify the general conclusions reached earlier although certain differences do apply. Both {1} and {2} below held in Section I, and, as shown in Appendix 2, {1} and {2} also hold with differing plots, multiple land use variables, and with a change in the price of land use services due to the renewal.[9] Specifically, if (a) for any given price structure for land use services, an increase in investment in property $j$ does not lower the value of property $i$, $i, j = 1, \ldots, n$, $i \neq j$, and if there are marginal externalities at $c$ and $c'$; and (b) the prices of land services in the neighborhood do not fall by too great an amount due to renewal (see Appendix 2C), it follows that:[10]

[8] There may be one or more unstable competitive equilibria for the neighborhood (designated as point $u$ in Figure 2; see fn. 4). Government controls would be necessary to stabilize the system at $u$ (although each property owner maintains his property at the $u$ level given that all other owners do the same). If a stable system is sought for the neighborhood by a renewal authority, $u$ can be viewed as a state solution since stability requires permanent government intervention; $u$ will be treated in Section III as a state solution. On property value relationships for $u$, see fn. 10.

[9] The price impact of renewal depends on the number of slum units removed and of new units created by renewal, the quality of the new units and the elasticities of demand (own and cross) and supply of all types of housing in the metropolitan region. This is analyzed by James Sweeney and the author (1975).

[10] {1} and {2} hold regardless of the number of competitive equilibria. For {1}, $c'$ and $c$ can be interpreted as any two competitive equilibria for which $I_i^{c'} > I_i^c$, all $i$; this includes an unstable competitive equilibrium such as $u$ in Figure 2 and therefore if $I_i^{c'} > I_i^u > I_i^c$, all $i$, then $V_i(c') > V_i(u) > V_i(c)$, all $i$ (to see this, in the Appendix 2 proof, first let $c = u$ and then let $c' = u$). By {2}, investment is greater for some or all $i$ (see condi-

{1} $V_i$ is greater at $c'$ than at $c$, *all i*, where "higher" competitive equilibrium $c'$ is defined to involve greater investment for all $i$.

{2} Investment is greater at $o$ than at $c'$ for *some* plots $i$, i.e., $I_i^o > I_i^{c'}$. Furthermore, if it is also assumed that, for any given price structure for land use services, at equilibrium $c'$ incremental investment on plot $j$ raises the marginal private product of investment on plot $i$, any $i \neq j$, then investment is greater at $o$ than at $c'$ for *all i*, i.e., $I_i^o > I_i^{c'}$, all $i$.[11]

In Appendices 2A, B, and C, {1} and {2} are established for the case in which plots may differ and land use service prices may change due to renewal; a single land use variable $Q$ is assumed. Appendix 2D extends the results of 2A, B, and C to the case of many land use variables.

The general arguments here are similar to those of Section I. {1} states that all property owners benefit from a shift from $c$ to $c'$. A movement from $c$ to $c'$ could be effected by private agreement between property owners or by government incentives, with no controls needed after $c'$ is achieved. In contrast, to maintain $o$ with fragmented ownership of the neighborhood, permanent controls are necessary since property owners benefit by undermaintaining their own structures. Single ownership of the neighborhood would also produce a cooperative equilibrium. To sustain a state solution, permanent government controls compelling owners to remain at that equilibrium are necessary regardless of the number of property owners in the neighborhood.

Even if condition (a) or (b) above is violated and {1} and/or {2} does not hold, there will be a cooperative solution $o$ and may be a higher stable competitive equilibrium $c'$ (or more than one higher competitive equilibrium) to which the state can uplift the neighborhood. If {1} is violated, $c'$ provides lower $V_i$ but involves greater investment than does $c$.[12] In this case, with $c$ and $c'$ stable competitive equilibria, each owner maximizes his own net returns by maintaining his property at equilibrium $c$ quality level if surrounding properties are also at $c$ quality level and by maintaining his property at $c'$ quality level if surrounding properties are also at $c'$ level; this is so even though owners earn more if all properties are at $c$ than at $c'$.[13] Stability of $c$ and $c'$ implies that only a temporary government incentive (for example, a short-term subsidy or tax) is needed to uplift the neighborhood to $c'$ and thereafter to maintain $c'$, since $c'$ will be automatically sustained once it is reached; permanent controls are unnecessary. If {2} does not hold, $o$ will involve an equal or lower level of renewal than solution $c'$. As given in {1} and {2}, $c'$ may be preferred by a renewal agency over $o$ in that permanent controls (private or state) or a change to sole ownership of the renewal area would be required to maintain $o$, but only temporary incentives are needed to permanently establish $c'$. The various factors that are relevant in choosing a re-

tions under {2}) at $o$ than at *any* competitive equilibrium (including the competitive equilibrium with the highest level of investment and including any unstable competitive equilibrium such as $u$); to see that {2} applies to unstable equilibrium $u$, simply let $c' = u$ in the Appendix 2 proof.

[11] Note that the result in the second sentence of {2} does not require that the price structure for land use services remain unchanged; it simply requires that $\partial^2 V_i / \partial I_j \partial I_i > 0$ for any given price structure (see Appendix 2C). Similarly, condition (a) above only requires that $\partial V_i / \partial I_j > 0$ for any given price structure.

[12] Strictly, if {1} is violated it is possible that $V_i$ is the same at $c'$ and $c$ in which case owners are indifferent between the two points.

[13] All $V_i$ can be lower at $c'$ than at $c$ if a fall in the price of land use services due to renewal more than compensates for the improved externalities from the upgraded neighborhood, i.e., if condition (b) and consequently {1} do not hold. If owners anticipate lower $V_i$ at $c'$, they will not seek $c'$ willingly and must be encouraged to do so by the state, e.g., with a temporary subsidy or tax. But once a locally stable $c'$ is achieved, each owner maximizes his own profits by remaining at $c'$ given that all or most of his neighbors do likewise.

newal goal are discussed in Section III.

### III. Implications for Urban Renewal

In this section, the results of Sections I and II are used to analyze the two common renewal techniques: *rehabilitation*, which provides for structure upgrading rather than replacement, with land use forms and ownership left relatively undisturbed; and *redevelopment*, which involves large scale clearance, new land uses, and changed ownership patterns. It is shown that the choices of a method and a level of rehabilitation or redevelopment depend on the externalities generated by the renewal project.

#### A. *Rehabilitation*

The renewal agency has two general means by which it can effect renewal of a neighborhood. First, it can change the private "decision framework" of the neighborhood, and second, it can institute government incentives to encourage owners to renew. The private decision framework of a neighborhood is categorized as $D_1$, fragmented ownership with competition; $D_2$, fragmented ownership with temporary cooperative planning by owners and competition thereafter; $D_3$, fragmented ownership with permanent cooperative planning; and $D_4$, sole ownership of the neighborhood. Decision framework therefore refers both to ownership and planning organization.

Category $D_1$ characterizes the initial slum equilibrium. The state could effect $D_2$ by encouraging an agreement between owners to upgrade their structures to a particular level but with no provision for future maintenance. As explained earlier, assuming no permanent state maintenance controls, this fails to provide lasting improvement with a unique competitive equilibrium since the neighborhood will regress to a slum at $c$; but, it follows from Sections I and II that with multiple competitive equilibria, $D_2$ would lead to a *permanent* shift of the neighborhood to a higher competitive equilibrium at $c'$. Both $D_3$ and $D_4$ result in cooperative equilibrium $o$ (assuming no government controls preventing this outcome) since under either scheme the long-run objective is to maximize total neighborhood returns. Category $D_3$ involves permanently enforced private agreements between owners to maintain $o$. Such agreements are necessary since each owner maximizes his own returns by maintaining his property below the cooperative level as long as other owners remain at the cooperative level.

Government incentives to encourage owners to improve and maintain their properties are categorized as $G_1$, no incentives; $G_2$, temporary incentives; and $G_3$, permanent incentives. The incentive might involve a maintenance subsidy to owners, or a maintenance code with a penalty if violated. An incentive is temporary ($G_2$) if it requires owners to initially upgrade from above $Q^c$ in Figure 1 or 2, with no permanent obligation to maintain that higher level; for example, an upkeep code with a duration of one year is temporary.[14]

The renewal approaches ($(D, G)$ combinations) will now be evaluated in terms of *effectiveness* and *efficiency* in achieving each alternative renewal goal. The renewal goals considered are $c$ (no renewal), $o$, $s$, and with multiple equilibria, $c'$. A given $(D, G)$ is *effective* for a goal only if it *permanently* achieves that goal. Thus, $(D_4, G_2)$ would not be effective in producing $s$ since once the temporary government incentives under $G_2$ expire, the neighborhood's sole owner under $D_4$ would shift to $o$ (to maximum $V$). A given $(D, G)$ is *efficient* in achieving a given goal if it is effective *and* if there is no other effective $(D, G)$ that can always (in all situations) achieve that goal at a lower cost. Note that

[14] Permanent renewal generally means a renewal period of at least twenty or thirty years. A permanent renewal measure remains in effect for the entire renewal period whereas a temporary measure does not.

TABLE 1—EFFECTIVE REHABILITATION OR REDEVELOPMENT STRATEGIES WITH A UNIQUE COMPETITIVE EQUILIBRIUM

| | $G_1$ | $G_2$ | $G_3$ |
|---|---|---|---|
| $D_1$ | $c$ | $c$ | $c, o, s$ |
| $D_2$ | $c$ | $c$ | $c, o, s$ |
| $D_3$ | $o$ | $o$ | $c, o, s$ |
| $D_4$ | $o$ | $o$ | $c, o, s$ |

TABLE 2—EFFECTIVE REHABILITATION OR REDEVELOPMENT STRATEGIES WITH MULTIPLE COMPETITIVE EQUILIBRIA

| | $G_1$ | $G_2$ | $G_3$ |
|---|---|---|---|
| $D_1$ | $c$ | $c, c'$ | $c, c', o, s$ |
| $D_2$ | $c'$ | $c'$ | $c, c', o, s$ |
| $D_3$ | $o$ | $o$ | $c, c', o, s$ |
| $D_4$ | $o$ | $o$ | $c, c', o, s$ |

the $D$ need not be simultaneous with the $G$ in a $(D, G)$ policy; for example, to achieve $o$, $(D_4, G_2)$ may initially involve a temporary government incentive program to encourage owners to upgrade their properties $(G_2)$ followed by a transfer of the neighborhood to a sole owner who maintains the neighborhood at level $o$ $(D_4)$. It is assumed below that renewal does not cause a major decline in the price of land use services throughout the metropolitan region and that therefore $\{1\}$ and $\{2\}$ of Section II hold; relaxing this assumption would cause only a minor change in the results.[15]

1. *Effective Renewal Strategies*

Assuming first a unique competitive equilibrium, it follows from the above discussion that *temporary* measures alone (i.e., $(D_1, G_2)$, $(D_2, G_1)$, or $(D_2, G_2)$) uplifting the neighborhood above $c$ in Figure 1 will result in reversion to $c$. Therefore, only permanent controls—whether private ($D_3$ or $D_4$) or state ($G_3$)—can be effective in producing $o$. To produce $s$, only $G_3$ is effective since owners will never seek $s$ unless encouraged to do so by the state. In Table 1, the entries for any $(D, G)$ indicate the renewal goals ($c$, $o$, or $s$) that the $(D, G)$ can be *effective* in achieving. For each goal, several alternative $(D, G)$ are effective; for example, $s$ can be achieved regardless of $D$ as long as the appropriate incentive under $G_3$ is instituted. Similarly, under $D_3$ or $D_4$ equilibrium $o$ can be produced with $G_1$, $G_2$, or $G_3$. All renewal goals appear in each element of the $G_3$ column since the government can permanently compel (through taxes, subsidies, etc.) owners to achieve any level regardless of $D$.

Consider next the case of multiple competitive equilibria. Since the neighborhood will remain at high equilibrium $c'$ once it is achieved, its permanent attainment does not require permanent state or private measures. Only short-run government incentives or a short-run cooperative effort by property owners (who, as noted earlier, all gain from a movement from $c$ to $c'$) is necessary to achieve $c'$. Table 2 indicates the $(D, G)$ that will achieve each renewal goal assuming the presence of multiple competitive equilibria. Notice that temporary cooperative action by owners under $(D_2, G_1)$ or $(D_2, G_2)$ will result in $c'$ (rather than in $c$ as with a unique competitive equilibrium) because of the gain by all owners from such a shift. Also, as indicated in element $(D_1, G_2)$ in Table 2, short-run government incentives under $G_2$ can result in either $c$ or $c'$. With owners acting competitively $(D_1)$, the incentives must be sufficiently strong if the neighborhood is to rise to $c'$. Weak measures will produce $c$ in the long run.[16]

[15] If prices fall radically then the only change in results is that, as noted in Section II, temporary government incentives may be necessary to achieve $c'$. In this case, $c'$ is to be deleted from the $(D_2, G_1)$ cell in Tables 2 and 3.

[16] The temporary state incentive must be strong enough to elicit investment given the assumption by each owner under $D_1$ that his neighbors will retain their properties in their existing condition. Even if it were

TABLE 3—EFFICIENT REHABILITATION OR REDEVELOPMENT STRATEGIES

| | $G_1$ | $G_2$ | $G_3$ |
|---|---|---|---|
| $D_1$ | $c$ | $c'$ | $o, s$ |
| $D_2$ | $c'$ | $c'$ | $o, s$ |
| $D_3$ | $o$ | $o$ | $s$ |
| $D_4$ | $o$ | $o$ | $s$ |

2. *Efficient Renewal Strategies*

An effective $(D, G)$ is efficient for a particular goal if no other effective $(D, G)$ exists which in all cases is less costly in achieving that goal. There may be more than one efficient $(D, G)$ for a given goal, with one of those efficient $(D, G)$ being the least costly under some circumstances and another of those $(D, G)$ the least costly under different circumstances.

To determine the *efficient* $(D, G)$, the "program costs" (designated $C$) of implementing the $(D, G)$ include $C_d$, the costs of organizing and enforcing private cooperative land use agreements under $D_2$ and $D_3$ and of transferring property rights to a single owner under $D_4$; and $C_g$, the costs of implementing government incentives and insuring that owners achieve and maintain the structure standards set by the state. Thus, $C=C_d+C_g$. Program costs do not include the physical investment in structure improvements (included in $I_i$ in Sections I and II) since these costs are the same for any given structure quality (renewal goal) regardless of the $(D, G)$ employed. It is the program costs $C$ that distinguish the various $(D, G)$ in terms of renewal costs. Program costs as defined here exclude any "social costs" of increased state control or increased private wealth concentration; however, the inclusion of these costs reinforces the results presented here.[17]

For each goal, an effective $(D, G)$ from Tables 1 and 2 is inefficient if it *necessarily* involves greater costs than do one or more of the other $(D, G)$ in achieving that goal. The effective $(D, G)$ that are not inefficient are the efficient $(D, G)$ in Table 3. It is assumed below that the same $(D, G)$ is used and the same equilibrium ($c$, $c'$, $o$, or $s$) is sought for all land use variables and all plots; the effect of relaxing this assumption is discussed at the end of Section IIIA.3. The efficient $(D, G)$ are determined as follows:

$c$: The least cost method for maintaining $c$ is to do nothing: $(D_1, G_1)$. Entry $c$ appears in the $(D_1, G_1)$ cell in Table 3.

$c'$: With multiple competitive equilibria, any approach (for $c'$) in the $D_3$ row or $G_3$ column of Table 2 is inefficient since it involves the short-run task of uplifting the neighborhood to $c'$ as do the temporary measures, but in addition, it involves more costly *permanent* (and unnecessary) private ($D_3$) or government ($G_3$) measures. Only $(D_1, G_1)$, $(D_2, G_1)$, and $(D_2, G_2)$ are efficient. The choice of $(D_2, G_2)$ might minimize costs by providing an initial application of temporary government structure improvement incentives ($G_2$) followed by temporary cooperative action by property owners ($D_2$) once the program has gained momentum.

$o$: With permanent cooperative private planning ($D_3$) or sole ownership ($D_4$), equilibrium $o$ will arise and consequently permanent government measures ($G_3$) are not needed; $(D_3, G_3)$ and $(D_4, G_3)$ are therefore inefficient. Note that $(D_3, G_2)$ and $(D_4, G_2)$ cannot be classified as inefficient since costs might be minimized through short-run government effort to lift the

assumed that each owner anticipates that his neighbor will renew due to the government incentive; the incentive would have to be sufficient to raise investment to a "self-sustaining" level (a $\overline{Q}_i$ greater than that at point $u$ in Figure 2, if the Section I model applies).

[17] Government control increases with higher numbered columns ($G_1$ to $G_2$ to $G_3$) and private wealth concentration increases with higher numbered rows (higher $D$). Including these social costs in $C$ does not alter the Table 3 results since it accentuates the rise in $C$ as column or row numbers increase.

neighborhood to $o$ followed by private enforcement of $o$ over the long term; and $(D_2, G_3)$ might be cost minimizing by providing for a temporary cooperative effort by owners to shift to $o$ with the government enforcing $o$ thereafter. Of the effective strategies for $o$, only $(D_3, G_3)$ and $(D_4, G_3)$ are inefficient.

$s$: None of the $(D, G)$ effective in producing $s$ is inefficient. Even $(D_2, G_3)$ could be cost minimizing by providing some renewal through short-run cooperative action $(D_2)$ followed by implementation and enforcement of an even higher renewal level using permanent government incentives. Therefore, all cells in the $G_3$ column are efficient for $s$.

The $(D, G)$ which minimizes costs for any goal will be one of those indicated in Table 3. For example, to achieve $c'$, $(D_1, G_2)$, $(D_2, G_1)$, or $(D_2, G_2)$ will minimize $C$. Which efficient $(D, G)$ minimizes $C$ will depend upon circumstances; for example, upon the number and cooperativeness of property owners, the dollar costs of private and government enforcement measures, etc. Notice that a particular $(D, G)$ and goal does not imply a unique policy; for example, to reach $o$ under $(D_1, G_3)$, the permanent government controls might involve tax penalties, subsidies, various enforcement and inspection schemes, etc. Consequently, in order to determine the efficient $(D, G)$ which minimizes $C$ *for a particular goal* (signified $(D, G)^*$), two steps are necessary: (i) for each efficient $(D, G)$ (in Table 3), determine the specific policy that achieves the goal with minimum cost; (ii) determine the efficient $(D, G)$ (and its cost minimizing policy determined under (i)) that minimizes $C$ for that goal. For example, if the goal is $o$, under (i) the least cost method for $(D_1, G_3)$ might be a particular subsidy program; the least cost method for reaching $o$ would also be determined for each of efficient strategies $(D_2, G_3)$, $(D_3, G_1)$, $(D_3, G_2)$, $(D_4, G_1)$, and $(D_4, G_2)$. Under step (ii), these six costs would be compared and the least cost efficient $(D, G)$, $(D, G)^*$, would be identified. Thus, $(D, G)^*$ for goal $o$ would be $(D_1, G_3)$ if the subsidy program mentioned above were the least costly of all methods for producing $o$.[18]

### 3. *The Choice of a Renewal Goal*

The selection of a renewal plan will depend upon what is included in benefits and costs and how those benefits and costs are weighted. However, whether Paretian efficiency alone or a social welfare function is the criterion used, program costs $C$ will likely be viewed as outlays to be deducted (along with other costs) from the benefits of a given plan. It follows that only the least cost efficient $(D, G)$, the $(D, G)^*$, are relevant in appraising alternatives. To determine the preferred goal and method of renewal, the net benefits of each possible state solution and of the competitive and cooperative equilibria are estimated by deducting all costs (including $C$) from the related benefits, *assuming the use of the relevant* $(D, G)^*$ *for each goal.*[19] Since all (private and state) revenues and costs must enter the analysis, net benefits must take into account the impact on government receipts and expenses.[20] Furthermore, the costs of renewal include any negative effects on other areas if the slum

[18] The cost comparison of the Table 3 strategies may proceed quickly in some cases. For example, if there are very large setup costs for private $(D)$ or government $(G)$ renewal strategies, it may be clearly uneconomical to use both $D$ and $G$ renewal measures (rather than just one). For rehabilitation in this case, the efficient $(D, G)$ in Table 3 would be: $(D_1, G_1)$ for $c$; $(D_1, G_2)$ and $(D_2, G_1)$ for $c'$; $(D_1, G_3)$, $(D_3, G_1)$, and $(D_4, G_1)$ for $o$; and $(D_1, G_3)$ for $s$.

[19] The number of state solutions is infinite; however, in most actual situations a manageably few state solutions can be identified which are clearly most acceptable (are feasible and yield the greatest net benefits) and these are then analyzed more carefully to determine the best state solution.

[20] On government revenues and renewal, see Groberg, Rothenberg and Slayton. See Jack Hirshleifer (1965, 1966) and Kenneth Arrow and Robert Lind on the appropriate private and public sector discount rates.

relocates to other parts of the city. That goal (which may be no renewal, $c$) with the highest net benefits is chosen:[21]

Solution $o$ which maximizes $V$ would be Pareto optimal assuming the nonexistence of government, zero renewal program costs ($C$), a competitive market for land use services in the metropolitan area, no inefficiencies in other sectors, and no neighborhood "social externalities" (for example, disease, crime, aesthetic costs, or wealth maldistribution due to the slum) at $o$ at nonoptimal levels.[22] However, all of these conditions do not hold in actuality and, therefore, even with Pareto efficiency the standard, cooperative equilibrium $o$ may not be preferred.

It was assumed in the above discussion that the same type of equilibrium ($c$, $c'$, $o$, or $s$) and the same ($D$, $G$) would be applied to all plots and to all land use variables (for example, we would not place structure quality and landscaping at $s$ and refuse control at $c'$). The analysis is more complex if equilibria or ($D$, $G$) differ between plots or variables, but the results in Tables 1, 2, and 3 still hold for each plot and variable. Furthermore, the advantage of using the suggested method (comparing only efficient ($D$, $G$)) is significantly greater in this complex case since many more policy combinations are involved.[23]

### B. *Redevelopment*

The above analysis (including Tables 1, 2, and 3) also applies to redevelopment, which ordinarily involves new land uses and multiple land use variables (see Appendix 2D on multiple variables). Redevelopment may raise the neighborhood to a higher competitive equilibrium or to a cooperative or state solution. In contrast with rehabilitation, however, under redevelopment a higher competitive equilibrium or cooperative equilibrium for all land use variables may fail to adequately approximate the renewal agency's plan for the neighborhood. A state solution will generally be sought by permanently requiring (a $G_3$ strategy) a particular general plan for the neighborhood. It may be that the state plan under $G_3$ covers certain aspects of the neighborhood, for example,

[21] The timing of renewal may affect the choice of a renewal goal since $c$, $c'$, and $o$ may shift over time with changes in the demand and supply for land use services. Also, neighborhood quality may not be static over time even if a given type of equilibrium ($c$, $c'$, $o$, or $s$) is maintained; e.g., maintaining $o$ (maximum $V$) over time may require slow neighborhood deterioration for some period, say sixty years, rejuvenation at time 60, etc. However, as long as the assumptions of the present paper hold at a point in time, {1} and {2} of Section II also hold at that point in time. If neighborhood changes over a renewal period (for example, 50 years) for any given type of equilibrium are small or at least not dominant in relation to the differences between equilibria (e.g., in terms of quality) at any moment in time, equilibrium estimates at the time of renewal can be useful for long range planning. If shifts are rapid, however, the time path of each equilibrium will be required.

[22] Three points are relevant here. First, if there is government, then $V$ (net value of property owner returns) will generally not equal the net value of all (private *and government* (*taxpayers*)) returns and costs from the neighborhood. Second, if the renewal area is large enough that renewal can affect the price structure for land use services, then maximizing $V$ introduces the inefficiencies arising when marginal cost is set equal to a marginal revenue that is less than price. Third, maximizing $V$ may not Pareto optimally deal with "*social externalities*"; Pareto optimality requires that any parties (e.g., taxpayers) desirous of paying for the removal of the externalities do so (e.g., through renewal) if consequently no one else is worse off; on such Pareto optimal wealth transfers, see Kenneth Boulding and the author (1972b).

[23] With differing types of goals ($c$, $c'$, $o$, or $s$) or ($D$, $G$) among land use variables, the following interdependencies are important: (a) the benefits, program costs, and renewal construction costs of any solution ($c$, $c'$, $o$, or $s$) and ($D$, $G$) for a particular variable depends on the solution achieved by all other variables; (b) program costs associated with a given solution and given ($D$, $G$) for a particular variable also depend on the ($D$, $G$) applied to all other variables; and (c) the level achieved by a particular land use variable at $c'$ or $o$ (or at $c$ with redevelopment, which is discussed in Section IIIB) depends on the solution achieved by all other variables. These interdependencies necessitate the comparison of the benefits and costs of all combinations of goals ($c$, $c'$, $o$, and $s$), efficient ($D$, $G$) (from Table 3), and land use variables. The combination with the highest net benefits should be adopted. The number of alternative combinations would be even further expanded if the solution ($c$, $c'$, $o$, or $s$) or ($D$, $G$) for any particular land use variable can differ between plots, e.g., $c'$ on plot $i$ and $o$ on plot $j$ for land use variable $h$.

type of structures, landscaping, parking facilities, but allows nonstate solutions for other land use variables, for example, upkeep, type of lessee tenants, work hours of employees (affecting traffic congestion), etc. The problem for renewal planners is whether a nonstate solution for these latter variables is better than a state solution, and, if so, which solution and policy should be pursued.

To analyze this problem, assume $n$ plots in the neighborhood *after* renewal and $m$ land use variables for each plot, for example, type of structure, quality of structure, and landscaping; define these variables for plot $i$ as $(Q_{i1}, Q_{i2}, \ldots, Q_{im})$. Based on the costs and benefits of various alternatives, the renewal authority establishes a plan under which $(Q_{i1}, \ldots, Q_{if}) \equiv {}_fQ_i$ are fixed at levels $(Q_{i1}^s, \ldots, Q_{if}^s) \equiv {}_fQ_i^s$, each $i$. Vector ${}_fQ_i$ might include type of structure, structure height limit, or quantities of parking and landscaped areas; ${}_fQ_i^s$ may differ between plots. Assume that variables $(Q_{ig}, \ldots, Q_{im})$ are *not* necessarily fixed in the plan, where $(Q_{ig}, \ldots, Q_{im}) \equiv {}_mQ_i$. Vector ${}_mQ_i$ could include structure and landscape upkeep, refuse control, or noise and traffic congestion associated with the use of plot $i$. Points $c$, $c'$, $o$, and $s$ are solutions for the ${}_mQ_i$ *given that* ${}_fQ_i = {}_fQ_i^s$, $i = 1, \ldots, n$. Assume for the moment that the same type of solution ($c$, $c'$, $o$, or $s$) and same $(D, G)$ apply to all $i$ for all land use variables $g$ to $m$.

Competitive equilibria $c$ and $c'$, cooperative equilibrium $o$, and any state solution $s$ are points that the ${}_mQ_i$ (for all plots $i$) can achieve given that ${}_fQ_i = {}_fQ_i^s$, $i = 1, \ldots, n$. The results of Section II fully apply here. As with rehabilitation, $c$, $c'$, or $o$ for the ${}_mQ_i$ may not provide conditions sought by the renewal agency. For example, at $c$ or even at higher competitive equilibrium $c'$, landscaping might be neglected, refuse unremoved, and traffic highly congested due to a failure by competitive property owners to control these factors; similar conditions might prevail at $o$. On the other hand, at one or more of these points ($c$, $c'$, and $o$), conditions might be highly acceptable, i.e., involve well-maintained landscaping, good traffic control, etc. The situations prevailing at $c$, $c'$, and $o$ will depend on the particular circumstances.

The renewal agency has the same $(D, G)$ alternatives for ${}_fQ_i^s$ and ${}_mQ_i$ as exist for rehabilitation. Many property owners (developers) may compete $(D_1)$ or cooperatively plan on a temporary $(D_2)$ or permanent $(D_3)$ basis; or the neighborhood may be transferred to a sole owner $(D_4)$. Government upkeep incentives may not be instituted $(G_1)$ or instituted temporarily $(G_2)$ or permanently $(G_3)$. Tables 1, 2, and 3 still fully apply. Note that the mode of ownership (fragmented under $D_1$, $D_2$, or $D_3$; or single under $D_4$) must be the same for ${}_fQ_i^s$ and ${}_mQ_i$ although, with fragmented ownership, it is possible in general to have different $D$ (of the first three) for different variables in ${}_mQ_i$ ($Q_{ig}$ to $Q_{im}$), for example, temporary cooperative planning $(D_2)$ for trash removal but permanent cooperative planning $(D_3)$ for structure upkeep; Tables 1 to 3 still apply to each land use variable.

As an example, to achieve $c'$ using $(D_1, G_2)$ the renewal authority could require that owners provide for trash removal during the first year after the development's completion. If each property owner maximizes profit by providing his own trash removal *given* that his neighbors do likewise, the state requirement can expire after the year and all property owners will maintain their sanitation facilities (i.e., remain at $c'$). With no such temporary government stipulation, all owners might initially fail to provide cleanup on the assumption that their neighbors planned to do the same, producing a competitive equilibrium $c$ at which all fail to provide for refuse removal.

As with rehabilitation, the benefits and costs (including $C$) of each potential goal ($c$, $c'$, $o$, or $s$) for ${}_mQ_i$ and its least cost

$(D, G)$ must be evaluated, but with only the efficient methods of Table 3 being considered. A cooperative or competitive equilibrium may approximately achieve the goals of the renewal authority and with lower $C$ than under a state solution, for example, if pollution or traffic congestion were limited at either equilibrium.

The analysis and Tables 1, 2, and 3 apply even if the type of equilibrium or the $(D, G)$ differ between plots or between land use variables; the conclusions at the end of Section IIIA concerning this issue also hold here.

## IV. Summary and Conclusions

In addition to their analytical interest, the present results should be helpful in planning actual renewal programs. In determining a plan, only efficient methods need be evaluated. This can greatly reduce the cost of analysis since the efficient set for any goal is appreciably smaller than the total set of possible strategies. Further, because any method that is efficient is also effective, adoption of an approach that will likely result in redeterioration of the neighborhood is avoided. Indeed, as was pointed out earlier, unless multiple competitive equilibria exist for a neighborhood, a renewal program with fragmented property ownership that does not include permanent private or state controls will lead to postrenewal deterioration, i.e., it will be inefficient and ineffective.

As a closing point, although the above analysis was specifically directed to the problem of urban renewal, many of the concepts outlined here are also applicable to the general issue of externalities and public policy. The policy measures utilizing private arrangements between parties and sole owner planning or employing state controls are applicable to most, if not all, cases involving technological externalities. In addition, the selection of policy actions on the basis of effectiveness and efficiency as described here should be appropriate for a wide range of situations in addition to those concerned with urban renewal. The assumptions used in deriving the effective strategies of Tables 1 and 2 and in determining the efficient strategies shown in Table 3 apply regardless of the source of externalities. The approaches in Table 3 are therefore efficient for a variety of situations. These results should be useful since the principle that only efficient strategies need be considered in evaluating alternatives would seem to hold in most cases.

## Appendix

### 1. *Identical Plot Competitive and Cooperative Solutions*

The Section I assumption of identical functions for all plots is not sufficient to insure that $Q_i = Q_j$, all $i, j$, at $c$, $c'$, or $o$. However, with the additional assumptions noted below it is so that $Q_i = Q_j$, all $i, j$, at $c$, $c'$, and $o$. To see this, note from equation (2), $\partial V_i/\partial Q_i = 0$ at $c$ or $c'$ and, from equation (3), $\partial V/\partial Q_i = 0$ at $o$. Differentiating these terms respectively:

$$\text{(A1)} \qquad \frac{dQ_i^c}{d\overline{Q}_i} = -\frac{(\partial^2 V_i/\partial \overline{Q}_i \partial Q_i)}{(\partial^2 V_i/\partial Q_i^2)}$$

$$\text{(A2)} \qquad \frac{dQ_i^0}{d\overline{Q}_i} = -\frac{(\partial^2 V/\partial \overline{Q}_i \partial Q_i)}{(\partial^2 V/\partial Q_i^2)}$$

The denominators of (A1) and (A2) are negative by the strict concavity of $V_i$ and $V$ (see fn. 7). The numerators are nonnegative by assumption (see fn. 4) and therefore (A1) and (A2) are nonnegative, i.e., $dQ_i/d\overline{Q}_i \geq 0$, all $i$, in both (A1) and (A2).

To see that $Q_i = Q_j$, all $i, j$, if $dQ_i/d\overline{Q}_i \geq 0$, all $i$, assume an equilibrium $E_1$ where $Q_g = Q^-$, $Q_k = Q^+ > Q^-$, any $g$, $k = 1, \ldots, n$ and $Q_i = Q_i^1$, all $i \neq g, k$. Since all $V_i$ functions and $I_i$ functions are identical, a second equilibrium $E_2$ must exist at which $Q_g = Q^+$, $Q_k = Q^-$ (same $g$, $k$ as for $E_1$) and $Q_i = Q_i^1$. Therefore, from $E_1$ to $E_2$, $\Delta Q_g = Q^+ - Q^-$, $\Delta Q_k = Q^- - Q^+$, and $\Delta Q_i = 0$, $i \neq g, k$. But this implies:

$$\text{(A3)} \quad (\Delta Q_g/\Delta \overline{Q}_g)$$

$$= \Delta Q_g \Big/ \left(\Delta Q_k + \sum_{i \neq g,k} \Delta Q_i\right)$$
$$= (Q^+ - Q^-)/(Q^- - Q^+) < 0$$

Equation (A3) implies that $dQ_g/d\overline{Q}_g < 0$ for some levels of $Q_g$ and $\overline{Q}_g$ which is inconsistent with the above result that $dQ_g/d\overline{Q}_g \geq 0$. Identical arguments hold for $Q_k$. Therefore, no equilibrium with $Q_g \neq Q_k$, any $g$, $k$, can occur. Hence, $Q_i = Q_j$, all $i$, $j$.

### 2. *Solutions for the General Case*

In Sections A, B, and C below it is assumed that (A4) and (A5) will hold where specifically indicated; land use service price changes are introduced in Section C and multiple land use variables are introduced in Section D.

$$\text{(A4)} \qquad \frac{\partial V_i}{\partial Q_j} \geq 0, \qquad \text{all } i, j, i \neq j$$

with the inequality near $c$ and $c'$

$$\text{(A5)} \qquad \frac{\partial}{\partial Q_j}\left(\frac{\partial V_i}{\partial Q_i}\right) \geq 0, \qquad \text{all } i, j, i \neq j$$

with the inequality near $c$ and $c'$

Let $\hat{Q}_i$ equal $Q$ on all plots other than $i$; $\hat{Q}_i = (Q_1, Q_2, \ldots, Q_{i-1}, Q_{i+1}, \ldots, Q_n)$. Define the net returns to $i$ given that $Q_i = Q_i^x$ and $\hat{Q}_i = \hat{Q}_i^t$ as $V_i(Q_i^x, \hat{Q}_i^t)$ where

$$\text{(A6)} \quad V_i(Q_i^x, \hat{Q}_i^t) = \int_0^{Q_i^x} \left(\frac{\partial R_i}{\partial Q_i}\right)_{\hat{Q}_i^t} dQ_i - I_i^x$$

where

$$(\partial R_i/\partial Q_i)\Big|_{\hat{Q}_i^t}$$

is the change in $R_i$ with respect to $Q_i$ given that $\hat{Q}_i = \hat{Q}_i^t$, $I_i^x$ is $I_i$ given that $Q_i = Q_i^x$, and $V_i = R_i(Q_i, \hat{Q}_i) - I_i(Q_i)$, with $R_i$ and $I_i$ defined in the text; assume that $(\partial I_i/\partial Q_i) > 0$ and $(\partial^2 R_i/\partial Q_i^2) < 0$.

A. *Net returns are greater at $c'$ than at $c$ for all plots, i.e., $V_i(Q_i^{c'}, \hat{Q}_i^{c'}) > V_i(Q_i^c, \hat{Q}_i^c)$, all $i$.* Proof: $V_i(Q_i^{c'}, \hat{Q}_i^{c'}) > V_i(Q_i^c, \hat{Q}_i^{c'})$, all $i$ since $Q_i^{c'}$ maximizes $V_i$ if $\hat{Q}_i = \hat{Q}_i^{c'}$; $V_i(Q_i^c, \hat{Q}_i^{c'}) > V_i(Q_i^c, \hat{Q}_i^c)$ using (A4) since $\hat{Q}_j^{c'} > \hat{Q}_j^c$ ($c'$ is the higher equilibrium). Combining these relations, $V_i(Q_i^{c'}, \hat{Q}_i^{c'}) > V_i(Q_i^c, \hat{Q}_i^{c'}) > V_i(Q_i^c, \hat{Q}_i^c)$; thus, $V_i(Q_i^{c'}, \hat{Q}_i^{c'}) > V_i(Q_i^c, \hat{Q}_i^c)$.

B. *With* (A4), *$I_i$ at $o$ exceeds $I_i$ at $c'$ for some plots $i$, i.e., $I_i^o > I_i^{c'}$, some $i$. With* (A5) *as well as* (A4), *$I_i^o > I_i^{c'}$, all $i$.*

To prove the above, we will prove Propositions (i), (ii), and (iii):

(i) At any point $e$ at which $Q_j^e = Q_j^{c'}$ and $Q_k^e < Q_k^{c'}$, $j$, $k = 1, \ldots, n$, $j \neq k$, $V_i^e < V_i^{c'}$, all $i = 1, \ldots, n$, i.e., all values are less at $e$ than at $c'$. Therefore, point $e$ cannot be cooperative equilibrium $o$ ($V$ is maximized at $o$).

(ii) At any point $w$ at which $Q_j^w \geq Q_j^{c'}$, with the inequality for one to $(n-1)$ of the $j$, and $Q_k^w < Q_k^{c'}$, $j$, $k = 1, \ldots, n$, $j \neq k$, $V_i^w < V_i^z$, all $i$, for point $z$ at which $Q_j^z = Q_j^w$ and $Q_k^z = Q_k^{c'}$. Therefore, point $w$ cannot be point $o$.

(iii) At $c'$, a small increase in all $Q_i$ above $Q_i^{c'}$, $i = 1, \ldots, n$, raises $V$, where the increase in $V$ is greater the larger the number of $Q_i$ increased above $Q_i^{c'}$.

Proposition (i) alone implies that at $o$, $Q_i^o > Q_i^{c'}$, one or more $i$; (i) holds assuming (A4). With (A4) and (A5), (i), (ii), and (iii) hold, implying that $Q_i^o > Q_i^{c'}$, all $i$. This implies that $I_i^o > I_i^{c'}$, all $i$, since $(\partial I_i/\partial Q_i) > 0$.

PROOFS:

(i) At $e$, $Q_j^e = Q_j^{c'}$, $j = 1, \ldots, f$, $Q_k^e < Q_k^{c'}$, $k = f+1, \ldots, n$; therefore $\hat{Q}_j^e < \hat{Q}_j^{c'}$ and $\hat{Q}_k^e \leq \hat{Q}_k^{c'}$, where the inequality holds if there is more than one $k$ plot. For $j = 1, \ldots, f$, $V_j^e \equiv V_j(Q_j^e, \hat{Q}_j^e) = V_j(Q_j^{c'}, \hat{Q}_j^e) < V_j(Q_j^{c'}, \hat{Q}_j^{c'}) \equiv V_j^{c'}$, all $j$, where $Q_j^e = Q_j^{c'}$; the inequality holds applying (1′) since $\hat{Q}_j^e < \hat{Q}_j^{c'}$.

For $k = f+1, \ldots, n$, $V_k^e \equiv V_k(Q_k^e, \hat{Q}_k^e) \leq V_k(Q_k^e, \hat{Q}_k^{c'}) < V_k(Q_k^{c'}, \hat{Q}_k^{c'}) \equiv V_k^{c'}$, all $k$. The first inequality for $k$ holds using (A4) and the fact that $\hat{Q}_k^e \leq \hat{Q}_k^{c'}$; the second inequality holds since $Q_k^{c'}$ maximizes $V_k$ given that $\hat{Q}_k = \hat{Q}_k^{c'}$. Therefore $V_i^e < V_i^{c'}$, all $i = 1, \ldots, n$, and point $e$ cannot be $V$ maximizing point $o$.

(ii) Define: point $w$ at which $Q_j^w \geq Q_j^{c'}$, $j = 1, \ldots, f$, with the inequality for one to $(n-1)$ of the $j$, and $Q_k^w < Q_k^{c'}$, $k = f+1, \ldots, n$; point $z$ at which $Q_j^z = Q_j^w$, $j = 1, \ldots, f$, and $Q_k^z = Q_k^{c'}$, $k = f+1, \ldots, n$, where the $j$ and $k$ relate to the same plots for $w$ and $z$. It fol-

lows that $\hat{Q}_j^w < \hat{Q}_j^z$; and $\hat{Q}_k^w \leq \hat{Q}_k^z$, the inequality holding if there is more than one $k$ plot. It will be shown next that $V_i^z > V_i^w$, all $i$, and therefore $w$ cannot be $V$ maximizing point $o$.

For $j=1, \ldots, f$, it follows from (A4) that $V_j^z \equiv V_j(Q_j^z, \hat{Q}_j^z) > V_j(Q_j^w, \hat{Q}_j^w) \equiv V_j^w$, all $j$, since $Q_j^z = Q_j^w$ and $\hat{Q}_j^z > \hat{Q}_j^w$. For $k=f+1, \ldots, n$, note that $V_k$ is maximized at $Q_k = Q_k^{c'}$ if $\hat{Q}_k = \hat{Q}_k^{c'}$ and therefore for $Q_j^z = Q_j^{c'}$, $V_k(Q_k^z, \hat{Q}_k^{c'}) = V_k(Q_k^{c'}, \hat{Q}_k^{c'}) > V_k(Q_k^w, \hat{Q}_k^{c'})$, i.e., $V_k(Q_k^z, Q_k^{c'}) - V_k(Q_k^w, Q_k^{c'}) \equiv D_1 > 0$. But, using (A5), we know in (A6) that

$$(\partial R_k/\partial Q_k)\Big|_{\hat{Q}_k^z} > (\partial R_k/\partial Q_k)\Big|_{\hat{Q}_k^{c'}}$$

since $\hat{Q}_k^z > \hat{Q}_k^{c'}$; thus, using (A6) $V(Q_k^z, \hat{Q}_k^z) > V(Q_k^z, Q_k^{c'})$. It follows that $V_k(Q_k^z, \hat{Q}_k^z) - V_k(Q_k^w, \hat{Q}_k^z) > D_1 > 0$; that is $V_k(Q_k^z, \hat{Q}_k^z) > V_k(Q_k^w, \hat{Q}_k^z)$. Also, using (A4), since $\hat{Q}_k^z \geq \hat{Q}_k^w$, $V_k(Q_k^w, \hat{Q}_k^z) \geq V_k(Q_k^w, \hat{Q}_k^w)$. Thus, $V_k^z \equiv V_k(Q_k^z, \hat{Q}_k^z) > V_k(Q_k^w, \hat{Q}_k^z) \geq V_k(Q_k^w, \hat{Q}_k^w) \equiv V_k^w$, $k = f+1, \ldots, n$. Therefore $V_j^z > V_j^w$, $j=1, \ldots, f$, $V_k^z > V_k^w$, $k=f+1, \ldots, n$; that is, $V_i^z > V_i^w$, all $i=1, \ldots, n$, and, therefore, point $w$ cannot be $V$ maximizing point $o$.

(iii) Assume initially that $Q_i = Q_i^{c'}$, all $i$ and therefore, $(\partial V_i/\partial Q_i) = 0$, all $i$. Raise $Q_j$ to $Q_j^x$, $Q_j^x > Q_j^{c'}$, any $j=1, \ldots, n$. By (A4), this raises $V$, $V = \Sigma V_i$. For all $Q_k$, $k \neq j$, $\hat{Q}_k > \hat{Q}_k^{c'}$ with $Q_j = Q_j^x > Q_j^{c'}$ and by (A5), $\partial V_k/\partial Q_k > 0$; also, for all rises in any $Q_{k'}$, $k'$ = a particular $k$, there is no decline by (A4) in $V_i$, $i \neq k'$, and there is no decline by (A5) in any $\partial V_i/\partial Q_i$, any $i \neq k'$. Therefore, for some rise in all $Q_k$, $k \neq j$, $\partial V_k/\partial Q_k > 0$, and $\partial V/\partial Q_k > 0$. A rise in all $Q_i$ above $Q_i^{c'}$, all $i=1, \ldots, n$, therefore raises $V$.

C. The proofs above required only (A4) and (A5) and did not require the constancy of the price of land use services. To relate (A4) and (A5) to price changes due to renewal, note that the value of plot $i$ equals $V_i = V_i(Q_i, \hat{Q}_i, P_i)$ where $P_i$ is the price of land use services provided by property $i$. Thus, in (A4),

$$\text{(A7)} \qquad \frac{\partial V_i}{\partial Q_j} = \left(\frac{\partial V_i}{\partial Q_j}\right)_{P_i} + \left(\frac{\partial V_i}{\partial P_i}\right)\left(\frac{\partial P_i}{\partial Q_j}\right)$$

where $(\partial V_i/\partial Q_j)_{P_i}$ is the change in $V_i$ with respect to a change in $Q_j$ given $P_i$. With no price change, $\partial P_i/\partial Q_j = 0$ and, using (A7), (A4) is satisfied if $(\partial V_i/\partial Q_j) = (\partial V_i/\partial Q_j)_{P_i} \geq 0$. Near $c$ and $c'$, $\partial V_i/\partial Q_j > 0$ by (A4). With price changes, to satisfy (A4) (A7) must be positive; that is

$$\text{(A8)} \qquad \left(\frac{\partial P_i}{\partial Q_j}\right) \geq -\frac{(\partial V_i/\partial Q_j)_{P_i}}{(\partial V_i/\partial P_i)}$$

At $c$ and $c'$, $(\partial V_i/Q_j)_{P_i} > 0$ by the assumption that (A4) holds for a fixed $P$; $(\partial V_i/\partial P_i)$ is necessarily positive. Therefore, by (A8), if the change in $P_i$ is not too large a negative magnitude, (A4) is satisfied. To examine (A5),

$$\text{(A9)} \qquad \frac{\partial}{\partial Q_j}\left(\frac{\partial V_i}{\partial Q_i}\right) = \frac{\partial}{\partial Q_j}\left(\frac{\partial V_i}{\partial Q_i}\right)_{P_i} + \left\{\frac{\partial}{\partial P_i}\left(\frac{\partial V_i}{\partial Q_i}\right)\right\}\left(\frac{\partial P_i}{\partial Q_j}\right)$$

With no price changes, $(\partial P_i/\partial Q_j) = 0$ and (A9) is simply (A5) given $P_i$. Earlier, we assumed (A5) to hold for a given $P_i$. If prices do change, then rearranging (A9), (A5) is satisfied if

$$\text{(A10)} \qquad \left(\frac{\partial P_i}{\partial Q_j}\right) \geq -\frac{\dfrac{\partial}{\partial Q_j}\left(\dfrac{\partial V_i}{\partial Q_i}\right)_{P_i}}{\dfrac{\partial}{\partial P_i}\left(\dfrac{\partial V_i}{\partial Q_i}\right)}$$

where the numerator of the right-hand term of (A10) is (A5) for a given $P_i$; this term was assumed positive at $c$ and $c'$ for a fixed $P_i$. The denominator of (A10) is necessarily positive. Thus (A9) and therefore (A5) are positive if $(\partial P_i/\partial Q_j)$ is not too large a negative quantity.

D. The above results also hold with $m$ land use variables. Let vector $Q_i = (Q_{i1}, \ldots, Q_{im})$ where $Q_{ih}$ is land use characteristic $h$, for example, plot size, structure quality, etc. Let $I_i$ be investment on plot $i$ where vector $\hat{I}_i = (I_1, I_2, \ldots, I_{i-1}, I_{i+1}, \ldots, I_n)$. Thus, $\hat{I}_i^{c'} > \hat{I}_i^c$ if $I_j^{c'} > I_j^c$, all $j \neq i$. As before, $V_i = R_i(Q_i, \hat{Q}_i) - I_i$. Analogous to (A4) we assume that

$$\text{(A11)}\quad \frac{\partial V_i}{\partial I_j} = \sum_{h=1}^{m}\left(\frac{\partial V_i}{\partial Q_{jh}}\right)\left(\frac{\partial Q_{jh}}{\partial I_j}\right) \geq 0$$
$$\text{all } i, j, j \neq i; \quad h = 1, \ldots, m$$

and, analogous to (A4), to establish (ii) and (iii) below, we assume that

$$\text{(A12)}\quad \frac{\partial}{\partial I_j}\left(\frac{\partial V_i}{\partial I_i}\right) \geq 0, \qquad \text{for all } i, j, j \neq i$$

where the inequalities hold for changes in $Q_j$ near $Q_j^c$ and $Q_j^{c'}$. At a competitive equilibrium, $\partial V_i/\partial Q_{ih}=0$ and, at $o$, $\partial V/\partial Q_{ih}=0$, all $i=1,\ldots,n$, all $h=1,\ldots,m$, where $V=\Sigma V_i$. The results of Appendix 2A follow as before since $V_i(Q_i^{c'}, \hat{Q}_i^{c'}) = V_i(I_i^{c'}, \hat{I}_i^{c'}) > V_i(I_i^c, \hat{I}_i^{c'}) > V_i(I_i^c, \hat{I}_i^c) = V_i(Q_i^c, \hat{Q}_i^c)$, where, as earlier, the first inequality holds since $I_i^{c'}$ maximizes $V_i$ given that $\hat{I}_i = \hat{I}_i^{c'}$ and the second inequality holds using (A11) and the fact that $\hat{I}_i^{c'} > \hat{I}_i^c$.

The proof in Appendix 2B that $I_i^o > I_i^{c'}$, all $i$, is virtually the same here as before except that investment $I_i$ rather than $Q_i$ is used in the arguments. Only proof of (i) is shown here to illustrate the approach. To show (i), define point $e$ at which $I_j^e = I_j^{c'}$, $j=1,\ldots,f$, and $I_k^e < I_k^{c'}$, $k=f+1,\ldots,n$; therefore, $\hat{I}_j^e < \hat{I}_j^{c'}$ and $\hat{I}_k^e \leq \hat{I}_k^{c'}$, the inequality holding if there is more than one $k$ plot. Thus, for $j=1,\ldots,f$, $V_j^e \equiv V_j(I_j^e, \hat{I}_j^e) = V_j(I_j^{c'}, \hat{I}_j^e) < V_j(I_j^{c'}, \hat{I}_j^{c'}) \equiv V_j^{c'}$, $j=1,\ldots,f$, using (A11) and the fact that $\hat{I}_j^e < \hat{I}_j^{c'}$. For plots $k=f+1,\ldots,n$, $V_k^{c'} \equiv V_k(I_k^{c'}, \hat{I}_k^{c'}) > V_k(I_k^e, \hat{I}_k^{c'}) \geq V(I_k^e, \hat{I}_k^e)$, $k=f+1,\ldots,n$. Thus, $V_j^e < V_j^{c'}$, $j=1,\ldots,f$, and $V_k^e < V_k^{c'}$, $k=f+1,\ldots,n$; that is, $V_i^e < V_i^{c'}$, all $i=1,\ldots,n$.

Propositions (ii) and (iii) are established in an analogous manner, and it follows that assuming (A11) and (A12) investment at $o$ exceeds investment at any competitive equilibrium.

The discussion and conclusions regarding price changes are as in Appendix 2C above except that $I$ instead of $Q$ is used in the equations.

## REFERENCES


**C. Abrams**, *The City is the Frontier*, New York 1965.

**K. J. Arrow and R. C. Lind**, "Uncertainty and the Evaluation of Public Investment Decisions," *Amer. Econ. Rev.*, June 1970, *60*, 364–78.

**M. J. Bailey**, "Note on the Economics of Residential Zoning and Urban Renewal," *Land Econ.*, Aug. 1959, *35*, 288–92.

**K. E. Boulding**, "The Basis of Value Judgement in Economics," in Sidney Hock, ed., *Human Values and Economic Policy*, New York 1967, 55–72.

**J. M. Buchanan and W. C. Stubblebine**, "Externality," *Economica*, Nov. 1962, *29*, 371–84.

**R. H. Coase**, "The Problem of Social Cost," *J. Law Econ.*, Oct. 1960, *3*, 1–44.

**O. A. Davis and A. B. Whinston**, "The Economics of Urban Renewal," *J. Law Contemporary Probl.*, Winter 1961, *26*, 105–17.

——— and ———, "Externalities, Welfare and the Theory of Games," *J. Polit. Econ.*, June 1962, *70*, 241–62.

**H. Demsetz**, "The Exchange and Enforcement of Property Rights," *J. Law Econ.*, Oct. 1964, *7*, 11–26.

**R. P. Groberg**, "Urban Renewal Realistically Reappraised," *J. Law Contemporary Probl.*, Winter 1965, *30*, 212–29.

**J. Hirshleifer**, "Investment Decision Under Uncertainty: Choice Theoretic Approaches," *Quart. J. Econ.*, Nov. 1965, *79*, 509–36.

———, "Investment Decisions Under Uncertainty: Applications of the State Preference Approach," *Quart. J. Econ.*, May 1966, *80*, 252–77.

**E. J. Mishan**, *Welfare Economics*, New York 1964.

**S. C. Myers**, "Procedures for Capital Budgeting Under Uncertainty," *Ind. Manage. Rev.*, Spring 1968, *9*, 1–19.

**H. Nourse**, "The Economics of Urban Renewal," *Land Econ.*, Feb. 1966, *42*, 65–74.

**C. R. Plott**, "Externalities and Corrective Taxes," *Economica*, Feb. 1966, *33*, 84–87.

**J. Rothenberg**, *Economic Evaluation of Urban Renewal*, Washington 1967.

**L. D. Schall**, (1971a) "A Note on Externalities and Property Valuation," *J. Reg. Sci.*, Apr. 1971, *11*, 101–05.

———, (1971b) "Firm Financial Structure and Investment," *J. Finance Quant. Anal.*, June 1971, *6*, 925–42.

———, (1971c) "Technological Externalities and Resource Allocation," *J. Polit. Econ.*, Sept./Oct. 1971, *5*, 983–1001.

———, (1972a) "Asset Valuation, Firm Investment, and Firm Diversification," *J. Bus., Univ. Chicago*, Jan. 1972, *45*, 11–28.

———, (1972b) "Interdependent Utilities and Pareto Optimality," *Quart. J. Econ.*, Feb. 1972, *86*, 19–24.

———, "The Behavior of Chain Markets with An Application to Housing and Urban Renewal," unpublished manuscript, Univ. Washington 1975.

W. L. Slayton, "The Operation and Achievements of the Urban Renewal Program," in James O. Wilson, ed., *Urban Renewal*, Cambridge 1967.

**P. O. Steiner**, "Choosing Among Alternative Public Investments in the Water Resource Field," *Amer. Econ. Rev.*, Dec. 1959, *49*, 893–916.

J. L. Sweeney, "A Commodity Hierarchy Model of the Rental Housing Market," *J. Urban Econ.*, July 1974, *1*, 288–323.

S. Wellisz, "On External Diseconomies and Government," *Economica*, Nov. 1964, *31*, 345–62.

# Regional and Distributional Effects of a Negative Income Tax

By FREDRICK GOLLADAY AND ROBERT HAVEMAN*

Although income transfer policies have economic consequences similar to those of resource-using public expenditures, analyses of their impacts have been narrower in scope. Evaluation of the equity effects of such programs has been limited to an accounting of the primary incidence of transfers and corresponding tax liabilities.[1] Research on their impacts on resource allocation has concentrated on labor supply,[2] headship and family structure,[3] fertility,[4] and migration.[5] It is clear, however, that such programs will provoke a wide variety of other economic adjustments whose net effect on the structure of output and the distribution of income is not readily apparent. Understanding these ultimate effects is essential for designing efficient and equitable income transfer policy.

## I

This paper presents numerical estimates of the micro-economic effects of redistributing purchasing power through a uniform, national program of income supplementation for the poor. The model developed here is rich in descriptive detail; however, because of its complexity, it is not amenable to solution by analytic methods. Moreover the model does not presume rapid, costless readjustment to a new equilibrium. Instead, it explores the sequence of economic decisions which are likely to be made in response to a redistribution of income. It focuses primarily on demand phenomena; it is argued that because of the nature and size of the policy change, neither significant supply responses nor price changes are induced.

The model provides a framework for evaluating income transfer policies which encompasses a wide range of economic impacts not previously considered. The effect of such programs on patterns of consumption expenditure is analyzed, and, by simulating a sequence of responses to these changes, the implications for regional output (by industry), employment (by industry and by occupation), and earnings (by earnings class) are estimated. By focusing on the impact of policy on goods and factor markets, the extent to which behavioral responses and market adjustments induced by the policy tend to reinforce or offset the initial or first-round allocation and distributional effects is determined.

The paper begins with a heuristic statement of the model and a discussion of the differences between its structure and the more familiar Leontief formulation of impact analysis. Then the model is presented formally and its empirical implementation

* Economist, World Bank, and professor of economics and Fellow, Institute for Research on Poverty, University of Wisconsin-Madison, respectively. Kevin Hollenbeck, Economist, Mathematica-Washington, provided extensive assistance. The research was conducted under the sponsorship of the Institute for Research on Poverty, and was supported in part by funds granted to the Institute by the Department of Health, Education, and Welfare pursuant to the Economic Opportunity Act of 1964. We wish to acknowledge the helpful comments on an earlier version of the manuscript of John Bishop, Sheldon Danziger, Irwin Garfinkel, Robert Lampman, Stanley Masters, Eugene Smolensky, and Harold Watts. The opinions expressed are our own.

[1] For example, see Robert Lampman and Charles Schultze et al.

[2] See Glen Cain and Harold Watts, and Watts et al.

[3] See Marjorie Honig.

[4] See Cain.

[5] See John Kain and Robert Schaefer.

is described. Finally, some empirical results of the model applied to a negative income tax proposal are reported and analyzed. The relationship of the first-round incidence to the indirect distributional effects of the policy is emphasized, and the impacts of the policy on various regions, industries, and occupations are described.

## II

Public policy measures can be interpreted as exogenous shocks to the economy which induce complex adjustments both in individual and firm behavior, and in markets. Such economic consequences of public sector decisions have been recognized in both regional impact studies[6] and recent analyses of the impact of aggregate economic performance on the size distribution of income.[7] The analytic structure for the former type of study derives from the modified, open Leontief system. The latter type of analysis seeks to translate policy-induced changes in macro-economic variables into changes in the income distribution through reduced form models relating the responses of occupational or demographic groups to aggregate changes. Inherent in both formulations are strong and often unacknowledged assumptions regarding the adjustments required to achieve a new equilibrium. Typically, it is presumed that adjustment will occur rapidly, and that the question of interest is the nature of the new equilibrium.

While the present model employs elements of the Leontief framework, it highlights the short-term adjustment process generated by the policy measure. This is accomplished by characterizing the response to the policy initiative as a sequence of private economic decisions occurring in the short run. Through this emphasis on the sequential nature of induced decisions, the adjustments themselves can be evaluated and the interequilibrium states can be observed.

In a general equilibrium analysis, any disturbance to an initial position would be traced through the series of induced behavioral adjustments, market responses, and income impacts until a new equilibrium position is attained. The framework employed here focuses on adjustments occurring in the initial period only. First, the disbursement of benefits and the incidence of taxes associated with an assumed income redistribution program is traced to individual households classified by income, region, and other socioeconomic characteristics. Then, the induced changes in item expenditures are evaluated for families grouped by region. Such shifts in final demands will call forth increases in the outputs of some industries in some regions and decreases in the outputs of others. These production responses are portrayed by an open interregional input-output system; this presumes that production equilibrium will be achieved within the period of analysis. The demand for labor—sectorally and spatially distinguished—will change in response to induced changes in gross output (by industry and region). Here again a Leontief technology is postulated and it is assumed that markets fully adjust within the period of analysis. As a result of shifts in the demand for labor, earnings flows by earnings class and region of recipient will change and the distribution of earned income will be altered.[8]

## III

The model is composed of five submodels (referred to here as "modules") which portray economic processes from the incidence of the initial transfer to the ultimate impact on the distribution of earnings. These processes are: 1) the direct

[6] See for example, Wassily W. Leontief et al. and William H. Miernyk et al.

[7] See Charles Metcalf and Thad Mirer.

[8] The model could be operated dynamically by introducing a module which would update demographic and economic variables at the end of each period. The system could then be solved recursively.

tax and transfer allocation process; 2) the consumption expenditure process; 3) the sectoral gross output process; 4) the factor demand process; and 5) the regional earnings distribution process.

The *tax-transfer module* simulates the incidence of changes in tax and transfer policies. The eligibility rules and benefit (tax) schedules implied by the policy are applied to a national sample of households, containing detailed information on demographic characteristics, earnings, and nonearned income.[9] From this simulation, estimates of the first-round benefits and tax liabilities generated by the policy are obtained for households classified by race, region, family size, education of the head, and preprogram income level. These estimates presume no work effort,[10] migration, or family structure responses to the program.

Assume $y_s$ is a vector, each element of which is the 1973 income level of a household included in the 1971 Current Population Survey (*CPS*).[11] Each household is indexed by family size, place of residence, marital status, and all other sociodemographic characteristics which enter the eligibility and benefit rules of the transfer and taxation programs to be simulated. Assume $\Phi$ is a diagonal matrix portraying the eligibility rules and benefit (tax) schedules of the program in the form of taxation and/or benefit rates corresponding to each element of $y_s$; $\Phi_{ii}$ (the $i$th element along the diagonal) is the taxation (and/or benefit) rate corresponding to the $i$th element of $y_s$. By premultiplying $y_s$ by $\Phi$, an estimate of the program-induced change in income is obtained for each family in the sample. Call this vector $t_s$.

$$t_s = \Phi \cdot y_s \tag{1}$$

Let $\Omega$ be a diagonal matrix of weights, which when applied to $t_s$ yields national estimates of program induced income changes.

$$t_n = \Omega \cdot t_s \tag{2}$$

The summation of all the elements of $t_n$ yields the total national transfer cost (or tax yield), $t_n^0$, of the program. For the combined transfer and taxation programs analyzed in the study, $t_n^0 = 0$ by assumption.

These estimates of policy-induced changes in family disposable income imply alterations in the level and composition of family consumption expenditures. Such expenditure responses are estimated in the *consumption expenditure module* through the application of savings propensities and marginal budget shares defined on current income to the estimated changes in family income.[12]

In estimating the change in total household expenditures resulting from the policy, all households in the *CPS* file are ag-

[9] Data from the Current Population Survey (*CPS*) were employed in this module. The March 1971 Survey (containing 1970 information on 50,000 family units) was aged to allow for demographic changes, economic growth, and inflation through 1973. In addition, the 6.0 percent unemployment rate of 1970 was adjusted to 4.9 percent by randomly assigning unemployment and duration of unemployment experiences to groups identified by age, sex, occupation, and unemployment experience using the RIM model developed by the Urban Institute. See Nelson McClung, John Moeller, and Eduardo Siquel.

[10] Recent studies indicate negligible reductions in labor supply for male family heads and moderate work effort reductions by spouses in response to income transfers. See Watts et al.

[11] Uppercase Greek letters represent matrices and lowercase Roman letters represent column vectors.

[12] By employing estimates of changes in current income, this approach implicitly assumes that changes in current rather than normal or permanent income motivate consumption behavior. However, if the policy change represents a permanent alteration in the income generation process, the evaluation of normal income will be affected. As a result, families not eligible for benefits in any given year may well experience increments in their normal income because of the policy and, in response, alter their consumption behavior. Moreover, because of transitory shifts in income, the change in normal income may well be quite different than the change in current income for families eligible for benefits in any given year. Simulation estimates of consumption responses to changes in normal income have been developed and, while not presented here, are available from the authors on request. The primary empirical results from this normal income approach are not, in substance, different from those presented here.

gregated into six preprogram income classes, and marginal expenditure propensities specific to each class are applied to the policy-induced change in household disposable income.[13] The resulting change in household expenditure is allocated among individual commodity categories by multiplying the expenditure change by a vector of marginal budget shares specific to the income class of the household.[14] From this calculation, induced changes in expenditures by detailed commodity categories are obtained for each household. By aggregating these vectors over all the family units in a region, a vector of policy-induced changes in expenditures within a region for each of 79 commodities is obtained.

More formally, let $M$ be a diagonal matrix in which the $i$th element along the diagonal is the estimated marginal propensity to consume from current income for the $i$th household. The elements of $M$ can assume one of six values depending on the pretransfer income class to which the corresponding household belongs. By premultiplying $t_n$ from the transfer module by $M$, the household consumption expenditure induced by the tax-transfer program is estimated for each weighted observation. This vector is represented by $c$.

$$c = M \cdot t_n \tag{3}$$

Summing the elements of $c$ yields $c^0$, an estimate of total national personal consumption expenditure induced by the program.

The estimated changes in aggregate household expenditure by region are allocated among individual commodity categories by multiplying the estimated change in total expenditure for each of the income groups in any region by a vector of estimated marginal budget shares for the 79 commodity categories for that income class. First, the vector $c$, which is indexed by region and income, is aggregated over households to form a new vector $\hat{c}$ of aggregate expenditures by region and income class. This is accomplished by premultiplying $c$ by an operator matrix $\Sigma$.

$$\hat{c} = \Sigma \cdot c \tag{4}$$

The marginal budget shares estimated from empirical data are used to construct a block diagonal matrix $\beta$. Each block along the diagonal is a submatrix of marginal budget shares for a region. The $i$th row of a submatrix is the vector of budget shares for the $i$th income class for that region. By premultiplying $\hat{c}$ by $\beta$ we obtain $e$, the vector of commodity specific expenditures of each region.

$$e = \beta \cdot \hat{c} \tag{5}$$

where

$$\beta = \begin{bmatrix} \beta_1 & & & & O \\ & \beta_2 & & & \\ & & \beta_3 & & \\ & & & \ddots & \\ O & & & & \beta_r \end{bmatrix}$$

[13] The marginal propensities are based on analysis of the 1960–61 Survey of Consumer Expenditures reported by Irwin Friend and Ronald Jones. They represent current expenditure out of current income. In the lowest income class where approximately a third of the households have higher permanent incomes, the marginal propensities are in excess of unity. The marginal propensities were adjusted for price level changes between 1960–61 and 1973. For the six preprogram income classes (in 1960–61 dollars), the marginal propensities are; less than \$2,500, 1.09; \$2,500–\$3,500, 1.01; \$3,500–\$4,500, .89; \$4,500–\$6,250, .84; \$6,250–\$8,750, .70; \$8,750 or more, .57.

[14] The marginal budget share vectors are obtained from a piecewise loglinear regression analysis in which household expenditures on each of 56 commodities were regressed against family income, family size, education of head, home ownership, region, age of head, and urban-rural location. The regression model was chosen to accommodate nonlinearities observed in previous studies and was patterned after that developed in John Moeller. The analysis employed household data from the 1960–61 Survey of Consumer Expenditure file containing commodity expenditure estimates for 343 categories and 14,000 consumer units.

$$(6) \qquad \beta_i = \begin{bmatrix} \overset{a}{b_1} & \overset{b}{b_1} \dots & \overset{n}{b_1} \\ \overset{a}{b_2} & & \\ \vdots & & \\ \overset{a}{b_6} & \dots\dots & \overset{n}{b_6} \end{bmatrix}$$

and in which $\overset{h}{b_g}$ is the marginal budget share for commodity $h$ and income class $g$. The marginal budget shares are obtained from marginal propensities to spend by commodities and hence sum to unity. It is assumed that $\beta_1=\beta_2= \dots =\beta_r$.

Policy-induced changes in the commodity and regional composition of consumption expenditures affect the demand for outputs of individual firms and result in a revised pattern of production throughout the economy. The ultimate impact of these changes on the structure of production is obscured by the complexity of interindustry and interregional dependencies. In the *regional gross output module*, the indirect regional and industrial production responses to changes in commodity expenditures are estimated.

The empirical module is adaptated from the Multi-Regional Input-Output Model (*MRIO*).[15] This model relies on a Leontief production technology—implying linearity, additivity, and nonsubstitutability—for each of 79 sectors in each of 44 regions.[16] The input-output requirements for the several sectors of this model were estimated individually for each region.[17] Trading patterns were estimated by dividing shipments from each region to each region by total uses in the receiving region; the column sum of the coefficients is therefore unity.[18]

The product of a block diagonal matrix of input-output coefficients with a matrix of trade coefficients estimates the input by commodity and origin per unit of output by sector. Solution of the Leontief equation then reveals the gross output by region and sector required to satisfy the increment to final demand.

In the previous module, policy-induced consumption demands were estimated for each of 79 production sectors in each of 23 regions and summarized in an 1817 element vector $e$ (equation (5)). Let $\Gamma$ be a square matrix of dimension 1817 x 1817 composed of 79 x 79 diagonal submatrices. Each element $(\Gamma_i^{gh})$ describes the fraction of total uses of commodity $i$ in region $h$ that is imported from region $g$. Further define $\hat{A}$ to be a block diagonal matrix (1817 x 1817) with 23 square submatrices (79 x 79) of input-output coefficients along the diagonal describing the structure of production in each region.

$$(7) \qquad \underset{1817 \times 1817}{\hat{A}} = \begin{bmatrix} A^1 & & & \bigcirc \\ & A^2 & & \\ & & \ddots & \\ \bigcirc & & & A^{23} \end{bmatrix}$$

$$\underset{79 \times 79}{A^r} = \begin{bmatrix} \overset{r}{a}_{1,1} & \dots & \overset{r}{a}_{1,79} \\ \vdots & & \vdots \\ \overset{r}{a}_{79,1} & \dots & \overset{r}{a}_{79,79} \end{bmatrix}$$

$$\underset{1817 \times 1817}{\Gamma} = \begin{bmatrix} \hat{\Gamma}^{1,1} & \hat{\Gamma}^{1,2} & \dots & \hat{\Gamma}^{1,23} \\ \hat{\Gamma}^{2,1} & \hat{\Gamma}^{2,2} & \dots & \hat{\Gamma}^{2,23} \\ \vdots & \vdots & & \vdots \\ \hat{\Gamma}^{22,1} & \hat{\Gamma}^{21,2} & \dots & \hat{\Gamma}^{22,23} \\ I & I & \dots & I \end{bmatrix}$$

$$(8) \qquad \underset{79 \times 79}{\Gamma^{xy}} = \begin{bmatrix} \Gamma_1^{xy} & & & \bigcirc \\ & \Gamma_2^{xy} & & \\ & & \ddots & \\ \bigcirc & & & \Gamma_{79}^{xy} \end{bmatrix}$$

[15] This model is described in Karen Polenske (1975).

[16] In order to ensure conformability between the regional breakdown in the prior modules and this module, it was necessary to reduce the dimensionality of the regionalization from 44 to 23. Aggregation was in transactions and shipments form.

[17] The regional technical coefficients are based upon the 1963 input-output study for the United States and upon sector studies for agriculture, mining, and construction. See Polenske (1974).

[18] The interregional trade flows are from a study conducted by John Rodgers of Jack Fawcett Associates and are based upon sources for the standard public modes of transportation. See John M. Rodgers (1973).

Multiplying $e$ by the matrix of trade coefficients [$\Gamma$], the allocation of final demands by industry among the regions is obtained. The resulting vector is multiplied by the inverse of an identity matrix minus the block diagonal matrix of input-output coefficients, which has been premultiplied by [$\Gamma$]. This calculation yields an 1817 element vector of gross output [$x$].

$$x = (I - \Gamma\hat{A})^{-1}\Gamma e \tag{9}$$

Equation (9) yields an estimate of the change in gross output in each of 79 industries in each of 23 regions resulting from the policy change. (In the empirical calculation, power series expansion was employed in estimating indirect requirements rather than general solution by inversion.)

The final two modules trace policy-induced demands back to the household sector—first by developing estimates of the pattern of induced labor demands by occupation and then by deriving the implications of this pattern for the distribution of earned income. The *occupational employment module* provides estimates of the occupational distribution of new job opportunities and, hence, of the effect of the program on the employment prospects of workers of various skill levels. As suggested above, these indirect effects may either reinforce or offset the primary antipoverty objective of the program.

Employment effects are estimated in two steps. First, the total change in regional and sectoral employment is obtained by multiplying unit labor requirements[19] by the detailed incremental gross outputs from the previous module. Second, the estimates of total sectoral employment are distributed over 114 occupational categories by applying a matrix of the percentage distribution of employment by occupation for each industry[20] to the sectoral employment estimates.[21]

Following the above notation, total employment requirements by industry and region [$n$] are obtained by multiplying the 1817 element gross output vector [$x$] by the matrix of employment-output coefficients [$\Lambda$] (composed of 23 block diagonal 79 x 79 matrices).

$$n = \Lambda \cdot x \tag{10}$$

The occupational composition of the employment is determined by multiplying the vector of incremental employment by sector and region $n$ by the matrix of occupational composition ratios [$\Theta$]. The matrix [$\Theta$] is block diagonal with submatrices [$\Theta^r$] of dimension 114 x 79, giving the occupational composition of employment by sector and region. In this study, it is assumed that [$\Theta^1$] = [$\Theta^2$] = ... = [$\Theta^{23}$]. The increment to the demand for labor in each of the 114 occupations in each of the 23 regions is given by the vector $w$.

$$w = \Theta \cdot n \tag{11}$$

Finally, the implications of this induced pattern of occupational demands on the marginal size distribution of labor income is ascertained in the *regional earnings distribution module*. It is assumed that any earned income attributable to incremental employment in an occupation and region is distributed by earnings class as existing earned income is distributed over those

[19] The employment-output coefficients are based upon employment by sector and region reported in the *MRIO* project. See Rodgers (1972). Annual compound productivity growth rates computed for each sector for the 1947–63 period are employed in aging the 1963 coefficients to 1973.

[20] The estimates of occupational composition by sector are from a *BLS* study based on the 1970 Census of Population. For a description of the methodology employed, see U.S. Bureau of Labor Statistics. It is assumed that the occupational employment patterns in any sector are constant across regions.

[21] While this is obviously a highly simplified model of the labor market, its use in a short-run, intraequilibrium framework is justified. Adjustments to relative wages are likely to be modest within a one-year period, and the nonlinearities in the derived demands for managerial and professional workers are in part offset by their lower supply elasticity.

currently employed in an occupation and region.[22] The incremental labor demand in each occupation and region (estimated in the previous module) is multiplied by the relative frequency distribution of all employed members of the labor force by earnings class in the corresponding region and occupation.[23] By aggregating these distributions over occupations within a region (or the nation), the induced employment by earnings class for the region (or the nation) is obtained. The structure of these marginal distributions may then be compared to the preexisting earnings distributions to determine the effect of the policy on high skill-high wage relative to low skill-low wage workers. Through such comparisons the impact of the indirect demands of the policy change in reinforcing or offsetting the primary incidence of the program can be ascertained.

This final module can be stated formally as

$$f = \Psi \cdot w \tag{12}$$

where $\Psi$ is a block diagonal matrix constructed from the 23 regional relative frequency distributions[24] and $f$ is a vector of the distribution of jobs induced by the transfer policy by earnings class.

## IV

This sequential simulation model is applied to a well-known income transfer proposal—a Negative Income Tax (*NIT*).[25] The *NIT* guarantees a minimum income of $800 per adult and $400 per child; all families with children are eligible for benefits. Beyond an earnings level of $720, the ratio of marginal benefits to marginal earnings is −.67; a benefit reduction rate of .60 is applied to unearned income. The transfer proposal is assumed to be financed by a surtax on the federal personal income tax. In the simulation, the sum of benefit and tax flows is set at zero, implying no net effect on the public sector budget.

Of the many detailed estimates obtained from implementation of the model, two are presented here.[26] In this section, the initial effect of the policy on the distribution of income will be compared with its ultimate impact on the earnings distribution. In Section V, the initial and ultimate regional effects will be compared.

The aggregate 1973 benefit payments from the *NIT* are estimated to be $3,421 million. The aggregate tax cost for the program is set equal to aggregate benefit payments, implying a federal personal income surtax of 3.67 percent. For any family or group of families, the *net benefit* (cost) from the program is equal to the benefit payment received (referred to as *gross benefit*) less the additional tax liability incurred.

In Table 1, simulation results of the distribution of the net benefits (transfers less taxes) of the proposal are shown. While net transfers accrue to all income classes below $6,000, all income classes above $6,000 incur net costs. The extent of the initial income redistribution accomplished by the plan is indicated by the disparities between the lowest and highest income classes. Families with incomes in excess of $15,000 are estimated to incur income losses of $2.3 billion while households with incomes below $2,000 stand to gain approximately $1.8 billion. When the gains and losses from the program are spread

[22] This implicitly assumes that the preexisting distribution of earned income reflects the distribution of skill, ability, and work effort of the pool of workers available in an occupation in a region.

[23] The regional relative frequency distributions are from a special tabulation using 1-in-100 tapes of the 1970 U.S. Census of Population. A total of 114 occupational categories and 23 regions are employed. Fifteen earnings classes are distinguished.

[24] Each block of $\Psi$ is a 15 x 114 matrix representing a regional distribution.

[25] In the full study, variants of this *NIT* were analyzed as well as the Family Assistance Plan of the Nixon Administration.

[26] A more complete presentation of the results is included in a forthcoming monograph, a draft of which is available from the authors on request.

TABLE 1—DISTRIBUTION OF NET TRANSFERS OF A NEGATIVE INCOME TAX BY INCOME CLASS, 1973

| Income Class | Net Benefits ($ millions) | Net Benefits Per Family |
|---|---|---|
| less than $1,000 | $991.8 | $511.69 |
| $1,000–$2,000 | 824.0 | 225.08 |
| $2,000–$3,000 | 808.5 | 193.74 |
| $3,000–$4,000 | 514.7 | 145.58 |
| $4,000–$5,000 | 186.8 | 57.13 |
| $5,000–$6,000 | 25.4 | 8.12 |
| $6,000–$10,000 | −315.0 | −23.54 |
| $10,000–$15,000 | −721.4 | −47.90 |
| $15,000–$20,000 | −699.3 | −77.57 |
| $20,000 or more | −1615.4 | −179.89 |

over all families within an income group, the average net benefit for the lowest income groups is nearly $500 and the average net loss to the highest income groups is over $250.

This first-round redistribution from higher to lower income groups is reflected in a narrowing of the income distribution from before to after implementation of the program. Prior to implementation, the Gini coefficient for the distribution of current family income was estimated to be .435.[27] If it assumed that *NIT* and its surtax financing arrangement is put into effect, the Gini coefficient would fall to .426, a reduction of nearly .01.

Whether this first-round redistribution is a valid measure of the ultimate equity impact of the program depends upon the distributional implications of the induced, indirect effects. By tracing the induced demands generated by the policy through the several simulation modules, the distributional pattern of the induced earnings can be estimated. If incremental employment is concentrated at the bottom of the earnings distribution, the demand for low skill-low wage workers will be stimulated (relative to the preprogram pattern of labor demand) and the indirect effects of the program may be said to reinforce the primary distributional impact.[28]

For all regions and for the United States, it is estimated that the *NIT* will result in an increase in aggregate labor demand. This is due to the lower marginal propensity to consume of higher income groups (who are net cost bearers) relative to lower income groups (who are net beneficiaries). The aggregate increase in labor demand is estimated to be about 120,000 positions. As a result, the high wage-low wage comparisons to be described focus on relative incremental labor demands by earnings class. The benchmark used is the composition of labor demands by earnings class *prior to* the program.

The induced effect of the program on any skill or earnings class is measured by an impact indicator.[29] This indicator compares the incremental labor demand in a regional earnings class with the preexisting employment in that class. Within a region, those earnings classes with an indicator in excess of (less than) that of the region as a whole will experience an increase (decrease) in their share of regional employment. Similar comparisons can be made for any regional earnings class relative to the induced national labor demand effect.

In Table 2, the induced earnings distribution impact of the *NIT* is presented for 23 detailed regions, 4 Census regions, and for the entire United States. One pattern dominates. The lowest skill/earnings

[27] The procedure for estimating the Gini coefficient is from N.C. Kakwani and N. Podder. For the preprogram Gini coefficient, the standard error was .03; it was .017 for the post-*NIT* Gini coefficient.

[28] Data on induced employment refer to individual jobs, and are not readily matched with family units—the basis of the size distribution of incomes.

[29] The earnings class indicator is the ratio of induced labor demand in a regional earnings class to 1970 employment in that earnings class times .001. Symbolically,

$$S_i = \frac{Mi}{.001\ (E_i)}$$

where $S_i$ is the impact indicator for a regional earnings class, $M_i$ is the program induced change in employment in that earnings class, and $E_i$ is the total 1970 employment in that earnings class.

TABLE 2—EARNINGS CLASS IMPACT INDICATORS FOR NET TRANSFERS FROM A NEGATIVE INCOME TAX BY REGION

| Region | Earnings Class: Less than $4,000 | $4,000–$10,000 | $10,000–$20,000 | More than $20,000 | Regional Impact Indicator | Induced Labor Demand (000) |
|---|---|---|---|---|---|---|
| **Northeast** | .03[a] | .11 | .11[b] | .05 | .08 | 2.1 |
| 1) CT, MA, ME, NH, RI, VT | .29[a] | .36 | .46 | .48[a] | .36 | 2.0 |
| 2) NY | −.29 | −.30 | −.25[b] | −.42[b] | −.32 | −2.7 |
| 3) PA, NJ | .22[a] | .35 | .46[b] | .39 | .32 | 2.8 |
| **North Central** | 1.00[a] | 1.12 | 1.23 | 1.29[b] | 1.08 | 28.1 |
| 4) OH, MI | .45[a] | .67 | .84[b] | .61 | .55 | 5.8 |
| 5) IN, IL | .65[a] | .83 | 1.03[b] | 1.00 | .80 | 6.1 |
| 6) WI, MN | 1.17[a] | 1.34 | 1.54 | 1.69[b] | 1.31 | 5.0 |
| 7) IA, MO | 1.58[a] | 1.76 | 1.96 | 2.21[b] | 1.72 | 5.9 |
| 8) KA NB, ND, SD | 2.34[a] | 2.48 | 2.95 | 3.39[b] | 2.50 | 5.7 |
| **South** | 2.54[a] | 2.72 | 3.35[b] | 3.22 | 2.69 | 72.9 |
| 9) DE, DC, MD | .22[a] | .29 | .43[b] | .23 | .28 | .7 |
| 10) VA, WV | 1.12[a] | 1.44 | 1.54[b] | 1.27 | 1.27 | 3.3 |
| 11) NC | 2.30[a] | 2.39 | 3.87 | 4.32[b] | 2.49 | 5.8 |
| 12) SC | 3.58[a] | 4.93 | 7.25 | 11.82[b] | 4.97 | 5.8 |
| 13) GA | 1.10[a] | 1.22 | 1.89 | 1.72[b] | 1.24 | 2.5 |
| 14) FL | .78[a] | .92 | 1.08 | 1.15[b] | .84 | 2.4 |
| 15) KY, TN | 2.51[a] | 2.72 | 4.26 | 4.76[b] | 2.81 | 8.4 |
| 16) AL | 3.28[a] | 3.85 | 4.67 | 4.76[b] | 3.66 | 5.1 |
| 17) MS | 13.16 | 16.40 | 25.08[b] | 12.56[b] | 15.00 | 12.8 |
| 18) AR, OK | 2.07[a] | 2.34 | 3.11[b] | 3.05 | 2.29 | 4.4 |
| 19) LA | 6.76[a] | 7.66 | 9.19 | 10.88[b] | 7.47 | 10.2 |
| 20) TX | 2.15[a] | 2.43 | 3.01 | 3.11[b] | 2.37 | 11.6 |
| **West** | .99 | .99 | .98[a] | 1.13[b] | .99 | 15.7 |
| 21) AZ, CO, ID, NM, UT, NV, WY, MT, AK | 2.25[a] | 2.43 | 2.62 | 3.15[b] | 2.40 | 9.0 |
| 22) WA, OR, HI | 1.37[a] | 1.47 | 1.58 | 2.14[b] | 1.46 | 4.3 |
| 23) CA | .22[a] | .26 | .34[b] | .33 | .26 | 2.4 |
| **United States** | 1.32 | 1.28[a] | 1.35 | 1.37[b] | 1.30 | 118.8 |

[a] The lowest impact indicator in a region.
[b] The highest impact indicator in a region.

class (less than $4,000) has the lowest impact indicator in 21 of the 23 regions. In the remaining 2 regions, the impact indicator of the lowest earnings class lies below that of the region, even though it is not the lowest among the earnings classes. The opposite result is observed for higher skill classes. For the $10,000–$20,000 earnings class, the impact indicator is the highest among the earnings classes in 9 of the 23 regions, and in *no* region is the impact indicator for this class below that for the region as a whole. The results for the highest skill level (more than $20,000) show a similar pattern: for 14 of the 23 regions the impact indicator for this earnings class exceeds that of any other earnings class. Indeed, if the two highest earnings classes are aggregated, the lowest earnings class would have the lowest impact indicator and the highest class would have the highest in all of the regions. A similar pattern is present when the impacts are simulated for the nation as a whole: the lowest impact is on the $4,000–$10,000 class and the highest relative impact is recorded for the above $20,000 earnings class.

The underlying economic adjustments

which account for the program's adverse effect on the employment of low skill-low wage workers can be identified by examining the policy-induced demands placed on detailed industrial and occupational categories. In a number of prominent low wage occupations incremental labor demands are very small or negative, while some recognized high wage occupations are relatively heavily impacted. The low skill-low wage occupations with very low or negative demand effects (together with their impact indicators)[30] include: Textile Operatives (−.18), Health Service Workers (.98), Other Service Workers (.78), Protective Service Workers (.55), and Farmers and Farm Workers (.95).[31]

These comparisons suggest the following: the final distributional effect of explicitly redistributional policies is likely to be weaker than that indicated by the target efficiency of the *net* transfers. While the induced consumption and production decisions would be expected to be less propoor than the initial redistribution, these induced effects, in fact, tend to undo in part the initial redistribution. The earnings increments are somewhat more heavily concentrated among high earnings classes than even the *preprogram distribution of earned income*. Low income families tend to spend their income increments on goods and services produced by relatively high earnings groups, while higher income families tend to concentrate their spending reductions on sectors employing workers with relatively low earnings. While the program will achieve some reduction in inequality, the indirect effects tend to shift the structure of employment away from low skill/low earnings classes and toward high skill/low earnings classes.

## V

When the indirect effects are analyzed from a regional perspective, much the same sort of moderating effect is observed. While net transfers are heavily concentrated on certain regions, the induced industrial output and employment impacts tend to fall heavily on other regions.

Table 3 presents the distribution of *net* transfers by region for the *NIT*. On balance the South receives substantial net transfers—over $650 million. Each of the other three Census regions incur net costs. The average Mississippi family experiences an increase in disposable income of $203 while the average family in New York incurs a liability of $37. In terms of this first-round flow-of-funds effect, then, there is a substantial redistribution of disposable income from the richer to the poorer states in general, and from the North to the South, in particular.

In response to these changes in disposable income, consumption demand increases in some regions while expenditures in other regions decrease. Some of these expenditure changes increase demand for goods produced by local or regional industries, but other increments to demand affect businesses far from the site of the increased expenditure. This is especially notable after the second, third, and *n*th order interindustry demands are accounted for.

Table 3 also presents measures of the distribution of induced gross output. For the nation, gross output increases by about $3 billion. About 53 percent of this increase accrues to the South, while the North Central region receives almost a third. The Northeast and West each account for 5 and 11 percent, respectively. Impact indicators are also shown in the table.[32] While these results indicate a sub-

[30] These indicators are to be compared to the national indicator of 1.3 (Table 2).

[31] The reverse story could be told by observing the detailed occupations with simulated demand increases yielding high impact indicators. These occupations tend to be relatively high skill-high wage occupations. They include Managers, Officials, Administrators, and Skilled Workers.

[32] This indicator is analogous to that described in fn. 29.

TABLE 3—DISTRIBUTION OF NET TRANSFERS OF A NEGATIVE INCOME TAX AND INDUCED PRODUCTION BY REGION, 1973

| Region | Total Net Transfers ($ millions) | Per Family Net Transfer ($) | Ratio of Gross Transfers to Taxes | Gross Output (million $) |
|---|---|---|---|---|
| **Northeast** | **$—415.5** | **$—26** | **.55** | **150.1** |
| 1) CT, ME, MA, NH, RI, VT | −61.0 | −16 | .70 | 48.6 |
| 2) NY | −222.1 | −37 | .40 | −24.8 |
| 3) PA, NJ | −132.4 | −21 | .62 | 126.3 |
| **North Central** | **—173.1** | **—9** | **.82** | **905.0** |
| 4) OH, MI | −145.2 | −23 | .60 | 265.3 |
| 5) IN, IL | −88.8 | −16 | .70 | 210.5 |
| 6) WI, MN | −11.6 | −4 | .92 | 131.9 |
| 7) IA, MO | 26.0 | 12 | 1.29 | 169.9 |
| 8) KS, NB, ND, SD | 46.5 | 32 | 1.78 | 127.2 |
| **South** | **655.0** | **33** | **1.73** | **1570.9** |
| 9) DE, DC, MD | −35.0 | −20 | .68 | 43.3 |
| 10) VA, WV | 7.2 | 4 | 1.08 | 78.4 |
| 11) NC | 52.4 | 33 | 1.83 | 134.0 |
| 12) SC | 66.5 | 77 | 2.90 | 83.6 |
| 13) GA | 15.9 | 11 | 1.28 | 75.4 |
| 14) FL | −4.5 | −2 | .97 | 53.0 |
| 15) KY, TN | 75.4 | 34 | 1.92 | 186.5 |
| 16) AL | 58.2 | 58 | 2.31 | 91.7 |
| 17) MS | 165.5 | 203 | 9.17 | 189.8 |
| 18) AR, OK | 36.9 | 29 | 1.64 | 97.1 |
| 19) LA | 122.8 | 98 | 4.00 | 220.3 |
| 20) TX | 93.7 | 27 | 1.55 | 317.7 |
| **West** | **—66.2** | **—6** | **.90** | **342.1** |
| 21) AZ, CO, ID, NM, UT, NV, WY, MT, AK | 77.5 | 29 | 1.58 | 165.2 |
| 22) WA, OR, HI | 12.9 | 6 | 1.13 | 80.3 |
| 23) CA | −156.6 | −22 | .62 | 96.6 |
| **Total** | ≈0 | ≈0 | | 2968.1 |

stantial impact on the South, its share of total gross output is significantly smaller than its share of either net transfers or consumption expenditures. There is a substantial leakage of generated demands out of the region. The extent of this leakage is indicated in Table 4 which shows 1) the ratio of gross output to consumption expenditures induced by the program for each of the major regions and the nation, and 2) the difference between per capita net transfers and per capita gross output induced by the program. While the ratio of gross output to consumption expenditures is 1.9 for the nation, it is between 1.4 and 1.5 for both the South and the West. For the North Central region, the ratio is 3.6. Similarly, while the North Central region experiences *negative* net transfers from the *NIT*, the difference between per capita gross output and per capita net transfers for that region is about 130 percent of that in the South and about 165 percent of that in the West. As a result of the structure of interregional and interindustry relationships, then, a substantial share of second, third, and *n*th order demands falls on sectors whose productive capacity is concentrated in the North Central region. The high concentration of net

TABLE 4—INDICATORS OF OUTPUT LEAKAGE FOR THE *NIT* SIMULATION

| Region | Ratio of Gross Output to Consumption Expenditures | Per Capita Gross Output less Per Capita Net Transfer ($) |
|---|---|---|
| **Northeast** | a | $13.57 |
| **North Central** | 3.6 | 19.01 |
| **South** | 1.4 | 14.60 |
| **West** | 1.5 | 11.69 |
| **United States** | 1.9 | 14.64 |

[a] This ratio cannot be calculated as estimated consumption expenditures are negative for this region.

transfers in low-income regions is eroded in the exchange and production process so that the stimulation of economic activity in these regions is substantially smaller than the pattern of net transfers would imply.

## VI

This study has extended the framework for evaluating the economic impacts of tax-transfer policy by the construction and implementation of a short-run, multiregional multisectoral model. The model permits analysis of both policy-stimulated resource reallocations and redistributional impacts; here the redistributional effects—by income class and region—were emphasized.

The analysis has demonstrated that the indirect output and employment effects of the *NIT* tend to offset its primary distributional impact. Although the proposal has as its primary objective poverty reduction, relatively more high skill-high wage jobs are created than low skill-low wage jobs. Moreover, although the bulk of net transfer benefits are targeted at low income (primarily southern) states, most of the indirect employment and output gains are recorded in the relatively prosperous North Central and Northeast regions.

While results such as these expand the basis for evaluating proposed transfer policies, they neither comprehend the full range of economic impacts nor are they immune from criticism. To facilitate empirical development of the model, several likely behavioral responses had to be suppressed; hence, induced changes in labor supply, migration, and household structure were not investigated. Consumption expenditure responses were based on changes in current disposable income; current account Leontief production functions implying homogeneity, linearity, and nonsubstitutability were employed; and constant market prices were presumed. No estimates were made of the distributional impacts of induced income streams other than earnings. It is expected (a priori) that the net impact of these omitted effects—especially those related to wage rate inflexibility,[33] restricted labor supply, absence of household labor and investment demands, and the distributional impact of nonlabor income—would tend to reinforce the indirect distributional impacts presented above. Moreover, the model is designed to evaluate short-term effects and does not illuminate long-run or comparitive equilibrium effects of policy changes.

For policy purposes, these results suggest that the regional and income redistributional impacts of direct tax-transfer policies may not be as powerful as evaluations based on only primary incidence effects indicate. Given the pattern of induced effects, such policies may lose some of their edge when compared to, say, public employment programs or other policies with high employment impacts on low-skill workers. Moreover, if the induced shift in the skill composition of labor demand is combined with some presumed policy-induced reduction in the supply of low-skill labor, the future budgetary costs of the program are likely to be in excess of the first-year estimates presented here.

[33] With imperfect occupational mobility, relatively full employment, or segmented labor markets, a shift in the structure of labor demands toward high skill occupations may lead to relative wage rate—and, hence, earnings—increases in these occupations. This result would magnify the adverse indirect distributional effects described above.

## REFERENCES

**G. Cain**, "The Effect of Income Maintenance Laws on Fertility in the United States," in *Population, Resources and the Environment*, Vol. VI, Washington 1972.

——— **and H. W. Watts**, *Income Maintenance and Labor Supply*, New York 1973.

**I. Friend and R. Jones**, *Conference on Consumption and Saving*, Philadelphia 1966.

**M. Honig**, "AFDC Income, Recipient Rates and Family Dissolution," *J. Hum. Resources*, Summer 1974, *9*, 302–22.

**J. Kain and R. Schafer**, "Income Maintenance, Migration, and Regional Growth," *Publ. Policy*, Spring 1972, *20*, 199–225.

**N. C. Kakwani and N. Podder**, "On the Estimate of Lorenz Curves from Grouped Data," *Int. Econ. Rev.*, June 1943, *14*, 278–92.

**R. J. Lampman**, "How Much Does the American System of Transfers Benefit the Poor," in L. H. Goodman, ed., *Economic Progress and Social Welfare*, New York 1966.

**W. Leontief et al.**, "The Economic Impact—Industrial and Regional—of an Arms Cut," *Rev. Econ. Statist.*, Aug. 1965, *47*, 217–41.

**N. McClung, J. Moeller, and E. Siquel**, "Transfer Income Program Simulator," working pap. 950–3, Urban Institute, Washington 1971.

**C. Metcalf**, "The Size Distribution of Personal Income During the Business Cycle," *Amer. Econ. Rev.*, Sept. 1969, *59*, 657–58.

**W. H. Miernyk et al.**, *Simulating Regional Economic Development*, Lexington 1970.

**T. Mirer**, "The Distributional Impact of the 1970 Recession," *Rev. Econ. Statist.*, May 1973, *55*, 214–24.

**J. F. Moeller**, "Household Budget Responses to Negative Income Tax Simulations," unpublished doctoral dissertation, Univ. Wisconsin, Madison, 1970.

**K. Polenske**, *The United States Multi-Regional Input-Output Model*, Lexington 1975.

———, *State Estimates of Technology, 1963*, Lexington 1974.

**J. M. Rodgers**, *State Estimates of Outputs, Employment and Payrolls, 1947, 1958, 1963*, Lexington 1972.

———, *State Estimates of Commodity Trade Flows, 1963*, Lexington 1973.

**C. L. Schultze et al.**, *Setting National Priorities: The 1974 Budget*, Washington 1973.

**H. W. Watts et al.**, "The Labor Supply Response of Husbands," *J. Hum. Resources*, Spring 1974, *9*, 181–200.

**U.S. Bureau of the Census**, *Current Population Survey*, Mar. 1971 (data tape).

———, *U.S. Census of Population, 1970* (data tape).

**U.S. Bureau of Labor Statistics**, *Tomorrow's Manpower Needs*, vol. IV, Bull. 1606, Washington 1969.

# Illiquidity, Consumer Durable Expenditure, and Monetary Policy

*By* FREDERIC S. MISHKIN*

In the literature on consumer durable expenditure,[1] monetary policy has a major impact either through interest rate[2] or liquid asset (real balance) effects. The theoretical justification for the inclusion of liquid assets as an important determinant of consumer durable expenditures is not particularly strong,[3] and results with this variable have been mixed.[4] Yet, even though there is a solid theoretical basis for monetary policy effects through interest rates, empirical econometric work has rarely found these effects to be substantial.[5]

One possible conclusion from research in this area is that monetary policy has only a marginal effect on consumer durable expenditures. Another possibility, however, is that channels of monetary policy as yet unexplored might be a crucial determinant of this type of expenditure.

This paper studies the neglected illiquid aspect of the consumer durable asset. It finds that increased consumer liabilities are a major deterrent to consumer durable purchases and increased financial asset holdings a powerful encouragement. The results show that monetary policy has a strong impact on consumer durable expenditure through two additional channels of monetary influence: 1) Monetary policy affects the price of assets in the economy. Consumer financial asset holdings, thereby affected, influence expenditure on durables. 2) Past monetary policy will have affected the cost and availability of credit, thus influencing the size of consumers' debt holdings and hence consumer durable expenditure.

The paper proceeds in the following way: the next section develops a model which determines the effects of consumer durable illiquidity on the desirability of this asset; the second section contains aggregate time-series tests of this model; and the final section discusses the implications for monetary policy and contains concluding remarks.

## I. Illiquidity of the Consumer Durable Asset

One aspect of the consumer durable asset that distinguishes it from financial assets is its illiquidity. Well-developed capital markets exist for most financial assets,

* Teaching assistant, department of economics, Massachusetts Institute of Technology. I want to thank Stanley Fischer, Paul Joskow, Franco Modigliani, William Poole, and the participants in the Money Workshop at M.I.T. for their helpful comments. I would also like to thank Roger Hankin for computational assistance. This research was conducted while I was the recipient of a NSF Graduate Fellowship. Further research support has been provided by the Federal Reserve Bank of Boston. The evidence and conclusions set forth here are solely my own, and do not indicate concurrence by the Federal Reserve System.

[1] See Franco Modigliani; Michael Hamburger; F. Thomas Juster and Paul Wachtel (1972a); Michael McCarthy; Albert Hirsch, Maurice Liebenberg, and George Green; Ta-Chung Liu and Erh-Cheng Hwa; and Otto Eckstein, Edward Green, and Allen Sinai.

[2] Classified with interest rate effects are the effects of installment credit terms.

[3] One justification for the inclusion of liquid assets in consumer expenditure equations is found in Arnold Zellner, David Huang, and L. C. Chau.

[4] In none of the models mentioned in fn. 1 does the liquid asset variable enter significantly and with the right sign—indeed it often enters with the wrong sign—in both equations for the autos and parts and nonauto components of consumer durable expenditure.

[5] Hamburger's study seems to be the only piece of empirical work where these effects are substantial. Yet, he only finds these powerful effects when interest rates enter his equations with very long lags.

and cash can be generated with a minimum of cost in time, money, and effort by selling them in their near perfect markets. Capital markets for used consumer durables are, on the other hand, highly imperfect. Durable goods are very heterogeneous, and much information which is costly to obtain is needed to determine their value.[6] Also the bulk and difficulty in handling of durables leads to high transaction costs in their purchase or sale. These transaction and information problems lead to a wide spread between the price the consumer receives from selling his used consumer durable and its value in use.[7,8]

A simple two-period model of the effects of consumer durable illiquidity on the desirability of this asset is developed below. It is shown that the nature of markets for consumer durables forces the consumer to take account of his balance sheet status, i.e., his debt and financial asset position, as well as the riskiness of his income stream, in determining the desired level of his consumer durables stock.

Assume that a consumer buys a unit of durables with price equal to unity at the beginning of period one. The durable's in-use value at the end of the period would be $1-d$, where $d$ is the depreciation rate.[9] Yet, if the consumer suffers a shortfall in income so that the durable good has to be sold in a distress manner, its full value cannot be realized. Its illiquidity stems from the imperfect nature of the used consumer durable capital market. The degree of this illiquidity will be described by the variable $q$ ($q<1$), which is the fraction of in-use value that can be realized from a distress sale. This formulation is quite general: it is not dependent on any specific type of illiquidity loss; it includes the loss from a low sales price as well as from transaction costs. If, as a result of an income shortfall, a distress sale of the durable at the end of the period is required to raise cash, then the realized value of the durable at the end of the period will be $q(1-d)$, where $q$ is less than one.

If there is no distress sale, the one-period opportunity cost $C_0$ of holding a durable rather than a financial asset will be:

$$(1) \qquad C_0 = 1 - (1 - d) + r = r + d$$

But if there is a distress sale as a result of an income shortfall, then:

$$(2) \qquad C_0 = 1 - q(1 - d) + r = r + d + (1 - q)(1 - d)$$

where[10] $C_0$= one-period opportunity cost

[6] For example, how well has the owner treated his durable, has it been damaged, how frequently has it been used, was it a lemon to start with, etc.

[7] The value-in-use is the present discounted value of the durable's flow of services.

[8] To see why costly information would lead to a spread between selling price and in-use value, see George Akerlof. In an extreme case no organized market might exist as a result of information problems. The absence of organized markets for many types of used consumer durable goods is quite common.

[9] In the case of a durable where there is a planned trade-in, the expected costs incurred in the trade-in—transactions and otherwise—are included in depreciation. The value of the durable at the end of the period reflects these costs.

[10] The opportunity cost in equation (2) assumes that a consumer cannot borrow to cover his income shortfall or that the cost of borrowing over and above the yield on financial assets is more than $(1-q)(1-d)$. It is well known that financial intermediaries are more than happy to make loans to consumers when they least need it and are extremely reluctant to make loans to consumers when they are in financial trouble. If the financial intermediary does make a loan at all to a consumer with an income shortfall, it charges a very substantial premium to compensate for the increased risk. Thus the assumption inherent in equation (2) is quite reasonable. If the difference between the borrowing cost and the yield on financial assets is less than $(1-q)(1-d)$, the consumer will borrow instead of selling his consumer durables. This can be incorporated into the above model by replacing $(1-q)(1-d)$ with the spread between the distress borrowing rate and the yield on financial assets. This leads to the same results as found in the text.

The reluctance of financial intermediaries to lend to consumers in financial trouble explains why most consumers hold debt and financial assets at the same time, even if borrowing costs for the consumer not suffering financial distress are somewhat higher than the yield on

of holding a durable, and $r$ = one-period return on financial assets (which is assumed certain).

We can now view the opportunity cost of holding durable goods in an uncertain world with a Tobin-Markowitz mean-variance framework. If the probability of making a distress sale is $p$ and not making a distress sale is $1-p$, then

$$(3)\qquad E(C_0) = p[r + d + (1 - q)(1 - d)] + (1 - p)(r + d) = r + d + p(1 - q)(1 - d)$$

$$(4)\qquad Var\,(C_0) = p(1 - p)[(1 - q)(1 - d)]^2$$

where $E$ and $Var$ are the expectation and variance operators, respectively.

A distress sale occurs whenever consumption[11] plus debt service (interest plus amortization) is larger than income, plus readily available financial assets; i.e., when

$$(5)\qquad DS + CON - Y - FIN > 0$$

where $DS$ = debt service
$CON$ = consumption
$Y$ = disposable income
$FIN$ = holdings of financial assets

The permanent income hypothesis implies that

$$(6)\qquad CON = k\overline{Y}$$

where $k$ = the propensity to consume out of permanent income, and $\overline{Y}$ = expected average (permanent) income. If income is a normally distributed random variable, then using the standard normal distribution formula we may write:

$$(7)\qquad p = f[(DS - FIN - (1 - k)\overline{Y})/\sigma_Y]$$

where $\sigma_Y$ = the square root of the income variance, with

$$\frac{\partial p}{\partial DS} > 0, \qquad \frac{\partial p}{\partial FIN} < 0$$

$$\frac{\partial p}{\partial \sigma_Y} > 0, \quad \text{and} \quad \frac{\partial p}{\partial \overline{Y}} < 0$$

since $k$ is usually assumed to be less than one.

Debt service is a positive function of the consumer's liabilities at the beginning of the period, hence

$$(8)\qquad \frac{\partial DS}{\partial DEBT} > 0$$

where $DEBT$ = liabilities at the beginning of the period. Now:

$$(9)\qquad \frac{\partial E(C_0)}{\partial DEBT} = (1-q)(1-d) \cdot \left[\frac{\partial p}{\partial DS}\,\frac{\partial DS}{\partial DEBT}\right] > 0$$

$$(10)\qquad \frac{\partial E(C_0)}{\partial FIN} = (1-q)(1-d)\left[\frac{\partial p}{\partial FIN}\right] < 0$$

$$(11)\qquad \frac{\partial E(C_0)}{\partial \sigma_Y} = (1-q)(1-d)\left[\frac{\partial p}{\partial \sigma_Y}\right] > 0$$

$$(12)\qquad \frac{\partial E(C_0)}{\partial \overline{Y}} = (1-q)(1-d)\left[\frac{\partial p}{\partial \overline{Y}}\right] < 0$$

and

$$(13)\qquad \frac{\partial\, Var\,(C_0)}{\partial DEBT} = [(1 - q)(1 - d)]^2 \cdot [1 - 2p]\left[\frac{\partial p}{\partial DS}\,\frac{\partial DS}{\partial DEBT}\right]$$

$$(14)\qquad \frac{\partial\, Var\,(C_0)}{\partial FIN} = [(1 - q)(1 - d)]^2 \cdot [1 - 2p]\left[\frac{\partial p}{\partial FIN}\right]$$

financial assets. When a consumer suffers a drop in income, financial assets are a buffer that help prevent the consumer from taking losses either by selling his durables or borrowing at inflated rates to raise cash; thus the consumer will not try to minimize his borrowings by holding no financial assets as he would in a world of absolute certainty and perfect capital markets.

[11] Since a distress sale can be avoided at a relatively low cost by a reduction in consumer durable expenditure, consumption, not consumer expenditure, is the relevant variable for the necessity of a distress sale.

$$(15)\quad \frac{\partial\, Var\,(C_0)}{\partial \sigma_Y} = [(1-q)(1-d)]^2 \cdot [1-2p]\left[\frac{\partial p}{\partial \sigma_Y}\right]$$

$$(16)\quad \frac{\partial\, Var\,(C_0)}{\partial \overline{Y}} = [(1-q)(1-d)]^2 \cdot [1-2p]\left[\frac{\partial p}{\partial \overline{Y}}\right]$$

If the probability of a distress sale is less than one-half ($p<1/2$) for consumer durables, which would certainly seem to be the case for most individuals in our economy, then[12]

$$(17)\quad \frac{\partial\, Var\,(C_0)}{\partial DEBT} > 0, \qquad \frac{\partial\, Var\,(C_0)}{\partial FIN} < 0,$$
$$\frac{\partial\, Var\,(C_0)}{\partial \sigma_Y} > 0, \qquad \frac{\partial\, Var\,(C_0)}{\partial \overline{Y}} < 0$$

In a Tobin-Markowitz mean-variance model, both a lower expected opportunity cost and a lower variance are preferred.[13] Therefore, a consumer durable is a more desirable asset: the lower the debt holdings, the higher the financial asset holdings, the lower the variance of income, and the higher is expected income in this period.[14]

## II. Time-Series Tests of the Liquidity Model

A stock adjustment model incorporating the results of the "liquidity" model of the previous section is developed here. It is tested on quarterly aggregate time-series data for consumer durables expenditure and its two component parts: autos and parts expenditure, and nonauto consumer durables expenditure. The models are estimated over the period 1954-I through 1972–IV, with the exclusion of quarters in which there were auto strikes, i.e., 1964-IV to 1965-II and 1970-IV to 1971-II.[15] All quantities are in real per capita terms (thousands of 1958 dollars per capita) with flows as seasonally adjusted annual rates.[16]

### A. *The Model*

The literature views a consumer durable as an asset in the portfolio which yields a return of consumption services; the consumer derives benefits from the services of the stock, not from the flow of durable purchases.[17] The consumer thus desires a

[12] As can be seen in an appendix available from the author, the assumption that $p$ is less than one-half is certainly not needed for the debt and financial asset results obtained here.

[13] If the consumer has a diversified portfolio, then the capital asset pricing model applies; he prefers a lower mean opportunity cost and a lower covariance with the market return. If the correlation of the opportunity cost of holding a durable and the market return is positive and reasonably constant, then a lower variance of the opportunity cost is preferred as in the simple mean-variance model used above. Richard Bower and Donald Lessard indicate that for most situations the simple mean-variance model usually leads to the same decisions as the capital asset pricing model.

[14] The model above is quite simple and gives a nice neat result, yet it does make the unrealistic assumption that consumption cannot be lowered below its desired level to meet the problem of an income shortfall, or that it would be more costly to do so than to incur a loss from distress selling a consumer durable. Furthermore, the mean-variance model used here requires special assumptions which have been objected to in the literature. A more general model, found in an appendix available from the author, has been developed which does not rely on the special assumptions of the mean-variance model and allows the consumer to meet an income shortfall by lowering his consumption below its desired level. The results for the effects of debt and financial asset holdings on the desirability of the consumer durable asset are the same in this model as in the mean-variance model presented above. The more general model is not used here because its exposition is not as simple, and because the role of income stream riskiness is not as clear.

[15] Strong strike effects are felt in both the quarter of the strike and the quarter following. Use of first-order serial correlation corrections necessitates excluding the second quarter following the strike from the sample period as well as the two previous quarters in the consumer durables and autos and parts estimations. These quarters were also excluded for the nonauto consumer durables estimations because aberrations in the auto sector might have an impact on nonauto durable purchases. In fact, model estimates for the nonauto consumer durable sector were not appreciably affected when the excluded quarters were included in estimating the models.

[16] The sources of these data are described in another appendix available from the author.

[17] See Arnold Harberger, Gregory Chow, Modigliani, Richard Stone and D. A. Rowe, and Juster and Wachtel (1972a).

certain stock of durables which is a function of permanent income and the user rental cost of capital. The liquidity model developed in the previous section indicates that, in addition, the desired durables stock is a function of the value of the consumer's debt and financial asset holdings at the beginning of the period. Therefore:

$$K^* = f(Y_P, CAPC, DEBT, FIN) + E_A \tag{18}$$

where $K^*$ = real per capita desired stock of durables,

$Y_p$ = real per capita expected average (permanent) income,

$CAPC$ = user rental cost of consumer durable capital[18] $(RCB+D)(PCD/PCON)$,

$RCB$ = Moody's AAA corporate bond rate,

$D$ = annual depreciation rate,[19]

$PCD$ = consumer durables implicit price deflator,

$PCON$ = consumption implicit price deflator,

$DEBT$ = real per capita debt holdings of households—beginning of quarter,

$FIN$ = real per capita gross financial asset holdings of households (includes demand deposits plus currency, time and savings deposits, bonds, corporate equity, life insurance and pension funds, and other miscellaneous assets)—beginning of quarter,

$E_A$ = additive error term.

When expected income is high, and the desired durables stock is high, a change in the user capital cost should cause a larger dollar change in the desired stock of durables. Thus, equation (18) is linearized with the coefficient of permanent income a linear function of the user rental capital cost,[20] i.e.,

$$K^* = a + (b + c\,CAPC)\,Y_P + d\,DEBT + e\,FIN + E_A \tag{19}$$

Consumer durable expenditure is modeled with the stock-adjustment or so-called flexible-accelerator model which views consumers as adjusting only slowly to their desired stock of durables. The change in the stock, i.e., net investment, is only a fraction, $\lambda$, of the gap between the desired and actual stock at the beginning of the period. Net investment is also viewed as a function of transitory income because: 1) some portion of transitory income and hence saving should be reflected in consumer durable purchases; and 2) transitory income is a proxy to some extent for perceptions of income variance[21,22] which

[18] The user rental cost of consumer durable capital used here is completely analogous to the user rental cost of capital in the investment studies of Robert Hall and Dale Jorgenson and of Charles Bischoff. The interest rate in the formula above is a nominal interest rate, not a real interest rate as would be appropriate in the Hall-Jorgenson formulation; thus the effect of inflation on consumer durable expenditure is not incorporated into this model. Attempts were made to estimate the effect of inflation on consumer durable expenditure and include it in the model, yet experiments with varied distributed lags of past inflation rates proved fruitless; no significant effects could be obtained. This is not surprising for the effect of inflation is by no means clear. On one hand, with constant nominal interest rates inflation lowers the user rental cost of capital and encourages durable expenditures. Yet evidence from consumer surveys indicates that inflation increases consumers' perceptions of uncertainty (see Juster and Wachtel, 1972b), and this has a depressing effect on consumer durable expenditures.

[19] The assumed depreciation rate used in calculating the capital cost measure for all consumer durables is .20, while it is .25 for autos and parts, and .15 for nonauto consumer durables.

[20] This assumption is not critical to our argument. If $K^*$ is alternatively assumed to be a linear function of the right-hand side variables in (18), i.e.,

$$K^* = a + bY_P + c\,CAPC + d\,DEBT + e\,FIN + E_A \tag{19a}$$

the fit of the estimated model and the asymptotic $t$-statistics of the coefficients (except for the constant term) change hardly at all, and the important empirical results of this paper still hold.

[21] Transitory income is a cyclical variable which is related to the probability of a worker losing his job and suffering an interruption of his normal income stream.

the liquidity model indicates affects the desired stock of durables and hence net investment.[23] Therefore:

$$(20)\quad (K - K_{-1}) = \lambda(K^* - K_{-1}) + fY_T + E_B$$

where

$K$ = real per capita stock of durables at the end of quarter,
$\lambda$ = the quarterly adjustment rate,
$Y_T$ = real transitory income per capita,
$E_B$ = additive error term,

and subscripts refer to the time period of the $K$ variable.

Consumer durable expenditures, or equivalently, gross investment in consumer durable goods, equals the sum of net investment and replacement. Assuming a quarterly replacement rate of $\delta$:

$$(21)\quad EXP/4 = \delta K_{-1} + (K - K_{-1})$$

where $EXP$ = real per capita consumer durable expenditures at an annual rate. Combining equations (18) through (21) we derive the model to be estimated:

$$(22)\quad EXP = 4\lambda a + [4\lambda b + 4\lambda c\, CAPC]Y_P + 4\lambda d\, DEBT + 4\lambda e\, FIN + 4fY_T + 4[\delta - \lambda]K_{-1} + u$$

where $u$ = additive error term $= 4(\lambda E_A + E_B)$.

The signs of all the coefficients of equation (22) are easily determined. The coefficients on permanent and transitory income should both be positive because increased permanent or transitory income encourages consumer durable purchases.[24] Increased user capital costs should discourage purchase of consumer durables; this implies that $4\lambda c$ is less than zero. The lagged stock coefficient will be negative if the speed of adjustment is higher than the replacement rate—the usual case.

The results of the previous section indicate that illiquidity of the consumer durable asset should lead to a positive $FIN$ coefficient and a negative $DEBT$ coefficient in the above model. Changes in the value of financial assets for the wealthy, for whom liquidity is not a problem, might have a smaller impact on consumer durable expenditure than for the middle or lower income groups. For this reason, the

When transitory income is low, workers have a high probabilty of being laid off and have a larger income variance, and when it is high, workers have a low probability of being laid off and have a correspondingly lower income variance.

[22] The unemployment rate is also a cyclical variable that reflects the probability of losing one's job and is related to income stream variance. If transitory income is excluded from the expenditure model and the unemployment rate is used as a proxy for income variance in its place, it enters with the appropriate negative sign (indicating that higher income variance depresses consumer durable demand). It is statistically significant at the 5 percent level or higher in regression models for all consumer durables and its two component parts: nonauto consumer durables and autos and parts. The debt and financial asset variables results are not qualitatively different when unemployment is used in the expenditure models instead of transitory income.

[23] Attempts to find further measures of perceived income variance were unsuccessful. The unemployment rate, the Survey Research Center (*SRC*) consumer sentiment index, a filtered version of this index (see Juster and Wachtel, 1972b), a crude measure of perceived risk in the financial markets using yield spreads between low grade corporate bonds and comparable government securities, and calculated income variance from past data, were all tested in the equation (22) model shown here. Only the unemployment rate and the filtered *SRC* index proved to be statistically significant in any regression equation. Both of these variables were significant in the autos and parts regressions, yet the transitory income and adjustment speed coefficient took on unreasonable values. Furthermore, both variables had the wrong sign in the nonautos regression. The failure to find further measures of consumers' perceptions of income variance is not a severe problem. The estimated effect of financial asset holdings on the desired consumer durables stock should in any case reflect perceived income variance effects because of high correlation of the perceived variance and asset measures. When perceived income variance increases, a higher risk premium would probably be used in discounting the earning streams of equity. This causes a lower valuation of equity; thus the value of financial assets falls. A strong negative correlation between the gross financial assets measure and perceived income variance is thus expected.

[24] The transitory income coefficient should be positive not only because transitory income might be saved in the form of consumer durables, but also because a rise in transitory income indicates that consumers' income variance may have declined, thus increasing the desired stock of durables and durable purchases.

unequal and highly skewed distribution of financial asset holdings in this country would tend to sharply lower the aggregate financial assets coefficient in a model estimated on aggregate time-series data. On the other hand, consumer liabilities are distributed far more equally than financial assets; thus the coefficient on consumer liabilities should still retain a high value in time-series estimations. Even though the liquidity model does not imply that for an individual the debt coefficient should be markedly larger in absolute value than the financial assets coefficient, this result might be expected in time-series estimates of these coefficients which reflect the distribution effects described above.

### B. *Empirical Estimates*

Equation (22)—whether it be estimated for expenditures on all consumer durables, or for autos and parts and nonauto consumer durables expenditures—is just one equation in a simultaneous system; thus simultaneous equation bias will result from ordinary least squares estimation. In the above model this bias would be especially severe for the debt coefficient.[25] To avoid least squares bias, an instrumental variable technique has been used.[26] Strong serial correlation is evident in all the regression equations, and to achieve efficient estimates a first-order serial correlation correction has been made using Ray Fair's method and the appropriate additional instruments.[27,28] The results for each sector are denoted by superscripts: $D$ for all consumer durables; $A$ for autos and parts; and $NA$ for nonauto consumer durables.

The estimates for consumer durables are as follows, with asymptotic $t$-statistics in parentheses. The coefficient on $u_{-1}$ is the first-order serial correlation coefficient.

$$(23)\quad \underset{}{EXP^D} = \underset{(-2.45)}{-.3378} + \underset{(3.89)}{.2693}\, Y_T + (\underset{(2.40)}{.4295} - \underset{(-2.41)}{.4527}\, CAPC^D) Y_P - \underset{(-.01)}{.0014}\, \overset{D}{K_{-1}} - \underset{(-4.63)}{.2167}\, DEBT + \underset{(4.08)}{.0453}\, FIN + .5527\, u_{-1}$$

$R^2 = .9932$; Durbin-Watson $= 1.90$; Standard Error $= .007529$.

The results are good. The coefficients of the debt and financial asset variables have the signs hypothesized by the liquidity model and are highly significant; the coefficients are over four times their respective asymptotic standard errors. The depressing effect of debt holdings on consumer durable purchases is quite substantial; for every \$1 of debt held at the beginning of the quarter, durable purchases at an annual rate will be decreased by 22¢. The value of financial asset holdings has a

[25] Ordinary least squares estimates of the debt coefficient would be severely biased upward if the error term is positively serially correlated—the usual case. A positive error last period would imply a positive error in the current period, while increased durable purchases last period—a result of the positive error term—would lead to increased debt holdings at the beginning of the current period. The debt variable and the error term would thus be positively correlated, and this would lead to an upwardly biased ordinary least squares coefficient estimate. A comparison of the ordinary least squares and instrumental variables estimates of equation (22) indicates that the bias in ordinary least squares estimates is of the predicted direction and is quite strong.

[26] The list of instruments includes unborrowed reserves at member banks plus currency outside of banks, the discount rate, exports, federal government expenditures, the effective rate of personal income tax, these five variables lagged one period, the constant term, and population.

[27] Except for the lagged stock coefficients, regression estimates where there was no correction for serial correlation were not appreciably different from the corrected regression estimates. The serial correlation corrected regressions exhibited a higher adjustment speed of desired to actual stocks.

[28] Ordinary least squares estimates using a Cochrane-Orcutt technique for autocorrelation correction are provided in an appendix available from the author. Qualitatively the results are similar to those in the text (i.e., signs and $t$-statistics), though coefficient estimates sometimes differ by as much as 30 percent.

significant positive effect on the demand for durables, though, as might be expected, it is not as strong as the depressing effect of debt; an extra dollar of financial assets held at the beginning of the quarter leads to $4\frac{1}{2}$¢ of increased durables purchases.

In addition, the $Y_T$, $Y_P$, and $CAPC^D$ coefficients are all significant and of the expected sign in the estimated equation above. The magnitudes of these coefficients are also quite reasonable; 27¢ of a $1 increase in transitory income is spent on consumer durables, while a $1 increase in permanent income leads to somewhere in the neighborhood of 34¢ of increased durables expenditures. At the means of the sample data the interest rate elasticity of consumer durables expenditure is $-.14$, while the price elasticity is $-.71$. The lagged stock coefficient implies that approximately 6 percent[29] of the discrepancy between desired and actual stocks of durables is made up within the quarter; this is an annual adjustment rate of 22 percent.

The consumer durables demand model presented so far only allows for lags in the adjustment of actual to desired consumer durable stocks; i.e., no decision lags are allowed in the consumer's determination of his desired stock. This assumption seems rather naive. The consumer may acquire information on his user rental cost of durables slowly, and thus his decision on his desired stock of durables may be influenced by past as well as present user rental costs. Capital gains or losses may not be considered fully part of financial assets until they are realized. Movements in common stock prices, which lead to unrealized capital gains or losses in the short run, should not have their full impact immediately; instead, the valuation of common stock would affect the desired consumer durables stock with a distributed lag.

To test for the possibility of the lags described above, experimentation with polynomial distributed lags of the user rental cost variable and stock market financial assets have been pursued. There is no improvement in the standard error of the regression or asymptotic $t$-statistics from a lag on the capital cost variable. It seems that the consumer does not take long to acquire information on his cost of capital. On the other hand, a substantial improvement in fit is obtained when the value of stock market assets affects the desired stock of durables with a distributed lag. The liquidity model implies that there should be no differences in the effect of stock market and non-stock market assets on consumer durable desirability; thus the sum of the lagged stock market asset coefficients should be equal to the coefficient of unlagged non-stock market financial assets. Applying this a priori equality as a constraint,[30] experiments with polynomial distributed lags constrained to be zero at the tail resulted in an endpoint constrained, second degree polynomial with a four-quarter lag having the best fit (lowest standard error of the regression). The result using instrumental variables and Fair's method is:

$$
\begin{aligned}
(24)\quad EXP^D = &\underset{(-3.30)}{-\,.5239} + \underset{(2.94)}{.2167}\, Y_T \\
&+ (\underset{(3.39)}{.7026} - \underset{(-3.25)}{.6409}\, CAPC^D) Y_P \\
&- \underset{(-1.18)}{.2630}\, K^D_{-1} - \underset{(-4.43)}{.3118}\, DEBT \\
&+ .0632\, NSFIN \\
&+ \underset{(3.50)}{.0231}\, STK + \underset{(4.61)}{.0173}\, STK_{-1}
\end{aligned}
$$

[29] This assumes a quarterly replacement rate of .05625, which is the depreciation rate used in computing the consumer durables stock.

[30] The null hypothesis that this constraint is valid cannot be rejected at the 5 percent level. This hypothesis was tested with a two-tailed asymptotic $t$-test. The asymptotic $t$-statistic equals .3276 while the critical $t$ at the 5 percent level is approximately two.

$$+ \underset{(2.75)}{.0121}\, STK_{-2}$$

$$+ \underset{(1.58)}{.0074}\, STK_{-3}$$

$$+ \underset{(1.03)}{.0034}\, STK_{-4} + .6383\, u_{-1}$$

$$\sum_{L=0}^{4} \text{coefficients of } STK_{-L} = \underset{(4.10)}{.0632}$$

$R^2 = .9940$; Durbin-Watson $= 2.01$; Standard Error $= .007104$.

$STK$ = real per capita value of households' stock market asset holdings—beginning of quarter.

$NSFIN$ = real per capita non-stock market financial asset holdings of households—beginning of quarter $= FIN - STK$.

The lag pattern of equation (24) has a desirable shape, with a stronger impact on durables expenditure from more recent movements of stock market asset holdings. Furthermore, the overall impact of gross financial consumer assets on durables expenditures is larger in the lagged equation (24), than in the unlagged version (23): the overall financial assets coefficient is .0632 in (24) vs .0453 in (23). The debt coefficient also increases in absolute value in (24); \$1 of increased debt holdings now leads to a 31¢ decrease in durables purchases. The $Y_T$ and $Y_P$ coefficients still have reasonable magnitudes in this regression and are significant at the 1 percent level, while the lagged stock coefficient now implies that over 12 percent of the discrepancy between desired and actual stocks is made up within the quarter—an annual adjustment rate of approximately 40 percent. This speed of adjustment is quite plausible and is in the middle of the range of estimated adjustment speeds in other consumer durable studies.[31]

[31] See Harberger.

A striking result of allowing a distributed lag on stock market financial assets is the increase in absolute value of the capital cost coefficient and the rise of its asymptotic $t$-statistic to a value over three. In this model, the user rental cost of capital, and hence interest rates, has a strong and significant effect on consumer durable purchases. At the sample means the interest rate elasticity of consumer durable purchases is $-.20$.

To put the regression results of (23) and (24) in perspective, it would be worthwhile to compare them to results from a regression which does not include the debt and financial asset terms which are implications of the liquidity model. Instrumental variable estimates using Fair's method for this "standard" stock-adjustment consumer durables model are as follows:

$$(25)\quad EXP^D = \underset{(-1.52)}{-.2205} + \underset{(2.01)}{.1954}\, Y_T$$

$$+ (\underset{(2.39)}{.4611} - \underset{(-3.10)}{.7982}\, CAPC^D) Y_P$$

$$- \underset{(-.23)}{.0535}\, K^D_{-1} + .7846\, u_{-1}$$

$R^2 = .9919$; Durbin-Watson $= 1.75$; Standard Error $= .008111$.

The regression results of equation (23) and especially (24), which incorporate the liquidity model, are much superior to the results of the standard regression (25). The fit is better and the autocorrelation coefficient—an indicator of specification error—is far lower. The $Y_T$ and $Y_P$ coefficients are not as statistically significant in the standard regression, and the speed of adjustment—a quarterly rate of 7 percent—is somewhat low.

The model of equation (22) has also been estimated for the autos and parts, and the nonauto consumer durables sectors separately. Regression estimates using instrumental variables and Fair's method

TABLE 1—AUTO AND PARTS REGRESSIONS

Instrumental Variables Estimates Using Fair's Method
Dependent Variable: $EXP^A$

| Coefficient of | Equations (26) | (27) | (28) |
|---|---|---|---|
| Constant Term | −.1920 | −.2591 | −.1570 |
| | (−2.89) | (−3.33) | (−2.14) |
| $Y_T$ | .1002 | .0777 | .0306 |
| | (1.42) | (1.08) | (.34) |
| $Y_P$ | .2142 | .3133 | .3432 |
| | (2.74) | (3.42) | (3.25) |
| $CAPC^A \cdot Y_P$ | −.1834 | −.2458 | −.4578 |
| | (−1.61) | (−2.09) | (−2.82) |
| $K^A_{-1}$ | −.0194 | −.2819 | −.4453 |
| | (−.08) | (−1.10) | (−1.40) |
| $DEBT$ | −.1731 | −.2149 | |
| | (−4.24) | (−3.63) | |
| $FIN$ | .0398 | | |
| | (4.09) | | |
| $NSFIN$ | | .0486 | |
| $STK$ | | .0209 | |
| | | (3.53) | |
| $STK_{-1}$ | | .0139 | |
| | | (4.05) | |
| $STK_{-2}$ | | .0083 | |
| | | (2.09) | |
| $STK_{-3}$ | | .0041 | |
| | | (.97) | |
| $STK_{-4}$ | | .0014 | |
| | | (.45) | |
| $\sum_{L=0}^{4} STK_{-L}$ | | .0486 | |
| | | (3.44) | |
| $\rho$ | .5163 | .6045 | .7630 |
| $R^2$ | .9703 | .9738 | .9661 |
| Durbin-Watson | 1.84 | 1.91 | 1.63 |
| Standard Error | .007138 | .006759 | .007513 |

*Note:* $\rho$ = First-order serial correlation coefficient. All other variables are as defined in the text. Asymptotic $t$-statistics in parentheses.

TABLE 2—NONAUTO CONSUMER DURABLE REGRESSIONS

Instrumental Variables Estimates Using Fair's Method
Dependent Variable: $EXP^{NA}$

| Coefficient of | Equation (29) | (30) | (31) |
|---|---|---|---|
| Constant Term | −.2697 | −.3220 | −.0753 |
| | (−3.31) | (−3.77) | (−1.51) |
| $Y_T$ | .1511 | .1327 | .1607 |
| | (6.26) | (5.16) | (5.90) |
| $Y_P$ | .3558 | .4291 | .1210 |
| | (3.32) | (3.83) | (1.99) |
| $CAPC^{NA} \cdot Y_P$ | −.2421 | −.3149 | −.2513 |
| | (−2.94) | (−3.58) | (−2.90) |
| $K^{NA}_{-1}$ | −.2633 | −.3644 | .1389 |
| | (−1.38) | (−1.85) | (1.08) |
| $DEBT$ | −.0672 | −.1021 | |
| | (−3.47) | (−3.74) | |
| $FIN$ | .0089 | | |
| | (2.52) | | |
| $NSFIN$ | | .0161 | |
| $STK$ | | .0046 | |
| | | (2.10) | |
| $STK_{-1}$ | | .0041 | |
| | | (3.17) | |
| $STK_{-2}$ | | .0034 | |
| | | (2.23) | |
| $STK_{-3}$ | | .0025 | |
| | | (1.55) | |
| $STK_{-4}$ | | .0014 | |
| | | (1.20) | |
| $\sum_{L=0}^{4} STK_{-L}$ | | .0161 | |
| | | (2.96) | |
| $\rho$ | .5758 | .5912 | .6325 |
| $R^2$ | .9974 | .9975 | .9970 |
| Durbin-Watson | 2.09 | 2.17 | 1.99 |
| Standard Error | .002574 | .002540 | .002724 |

appear in Tables 1 and 2. Experiments with endpoint constrained, polynomial distributed lags were also carried out for these sectors, and, as in the case for all consumer durable expenditures, the best fits were obtained with a four-quarter endpoint constrained, polynomial distributed lag on stock market assets. The constraint that the sum of the $STK$ coefficients should equal the coefficient on $NSFIN$ was imposed.[32] The estimates incorporating lags on stock market asset holdings also appear in Tables 1 and 2.

The results for both the autos and parts and nonauto consumer durable sectors are excellent. The debt and financial asset variables are of the right sign and are significant in all cases. The lag pattern of stock market assets in the lagged versions of the model is very similar in both sectors and has a sensible shape; more recent movements in the value of stock

[32] The null hypothesis that this constraint is valid cannot be rejected at the 5 percent level for either sector. The asymptotic $t$-statistic for the auto and parts and nonauto consumer durables regressions were .4294 and 1.1728, respectively. The critical $t$ at the 5 percent level is approximately two.

market asset holdings have greater impact on purchases, as in the estimates for all consumer durables. The $Y_T$, $Y_P$, capital cost, and lagged stock terms are all of the expected sign and are usually significant. The magnitudes of these coefficients are also reasonable. The lagged versions of the estimated model for the two sectors, equations (27) and (30), do have a superior fit to the unlagged models, (26) and (29); and the quarterly speed of adjustment implied by these lagged models is over 12 percent for autos and parts, and over 13 percent[33] for nonauto consumer durables; at annual rates these are both over 40 percent.

It is interesting to note that the estimated debt and financial assets coefficients are so much larger in the autos and parts regressions than in the nonauto regressions, in spite of the fact that autos and parts make up not quite half of total consumer durable purchases. The consumer's financial position seems to have more impact on his decision to purchase an automobile than it does on his decision to purchase a household consumer durable.[34] This is a worthwhile subject for further research.

Standard regression equations for both sectors where the debt and financial assets variables have been excluded have been estimated and appear as equations (28) and (31) in Tables 1 and 2. For both sectors, regression equations which incorporate the results of the liquidity model are superior to the standard regressions. They have a better fit, a lower standard error, and a smaller autocorrelation coefficient. Furthermore, in the nonauto consumer durables case the standard regression has an impossibly low speed of adjustment; only .5 percent of the discrepancy between desired and actual stocks is made up within the quarter—an annual rate of 2 percent.

Disaggregation of the consumer durable sector into its autos and parts and nonauto consumer durables components has resulted in further tests of the liquidity model. The results are still strongly supportive of this hypothesis.

## III. Implications for Monetary Policy and Concluding Remarks

The consumer durable expenditure model which incorporates the results of the liquidity model developed in this paper leaves monetary policy with a strong role to play in the demand for one of the most volatile components of gross national product. Three routes for monetary policy effects on consumer durable expenditures can now be envisioned.

1) Monetary policy affects interest rates and hence the user rental cost of capital. Tight monetary policy which raises interest rates will be a strong deterrent to consumer durable purchases because of the high interest elasticity of consumer durable demand indicated by empirical results in this paper.

2) In a Tobin, Foley-Sidrauski theoretical framework, monetary policy has a strong influence on asset prices in the economy. Tight monetary policy will lead to a fall in stock and bond prices and will thus result in a smaller valuation of the gross financial assets in the community. This will lead to decreased purchases of durables because consumers' financial positions have deteriorated; they are now left with a high probability of income shortfalls that would have to be met by the distress sale of consumer durables or a drop in consumption.

3) Past monetary policy will have affected the cost and availability of credit

[33] These adjustment rates assume a quarterly replacement rate of .07 for autos and parts and .045 for nonauto consumer durables.

[34] As a result of indivisibilities in the consumer's portfolio, the absolute size of the loss from selling a durable, and not just the loss per unit of the durable, could be important to consumer behavior. High priced durables such as automobiles would have a greater potential absolute loss from a forced sale than low priced durables, and this might explain the result found above.

to the consumer and will have thus affected the size of consumers' liabilities. Easy past monetary policy which has encouraged the buildup of consumer debt holdings will eventually prove a deterrent to future consumer durable purchases. The increased debt holdings force the consumer to desire more liquid assets.[35]

Viewing the consumer durable as an illiquid asset which must be traded in imperfect capital markets has led to a consumer durables demand model where perceived risk, and consumer liabilities and gross financial asset holdings influence consumer durables expenditure. In contrast to other work on macro-economic financial asset effects where net wealth influences consumer behavior,[36] this approach finds that the composition of the consumer balance sheet is critical to spending decisions.[37] The empirical estimates of this model have proved very encouraging, and several new and apparently potent channels of monetary policy that affect aggregate demand have been proposed. Furthermore, a traditional path for monetary policy effects on consumer durable expenditure has proved to be quite powerful in the model estimated here.

The liquidity model developed in this paper should also have applications in such areas as residential housing demand, and this will be the subject of further research. Many producer's goods, such as inventories and producer's durables are also illiquid assets; incorporating this feature into investment models might throw light on other possible channels of monetary policy effects in our economy. This avenue of monetary research should prove very fruitful.

[35] Simulations with a macro-econometric model (see the author) indicate that the two liquidity channels discussed in (2) and (3) far outweigh the interest rate effects discussed in (1), and are indeed extremely important in the determination of aggregate demand.

[36] For example, see Modigliani.

[37] An important implication of the analysis of this paper is that changes in the composition of the household balance sheet which leave net wealth unchanged can affect the expenditure behavior of households. An increase in indebtedness matched by an increase in holdings of nonfinancial assets which leaves net wealth constant would still lead to a future decline in consumer durable expenditure; a decrease in the value of financial asset holdings matched by an increase in nonfinancial asset holdings that left net wealth constant would also lead to a decline in consumer durable demand.

## REFERENCES


G. A. Akerlof, "The Market for 'Lemons': Qualitative Uncertainty and the Market Mechanisms," *Quart. J. Econ.*, Aug. 1970, *84*, 488–500.

C. Bischoff, "The Effects of Alternative Lag Distributions," in G. Fromm, ed., *Tax Incentives and Capital Spending*, Washington 1971, 61–130.

R. S. Bower and D. R. Lessard, "An Operational Approach to Risk Screening," *J. Finance*, May 1973, *28*, 321–37.

G. Chow, *The Demand for Automobiles in the United States: A Study in Consumer Durables*, Amsterdam 1957.

O. Eckstein, E. W. Green, and A. Sinai, "The Data Resources Model: Uses, Structure and Analysis of the U.S. Economy," *Int. Econ. Rev.*, Oct. 1974, *15*, 595–615.

R. C. Fair, "The Estimation of Simultaneous Equation Models with Lagged Endogenous Variables and First Order Serially Correlated Errors," *Econometrica*, May 1970, *38*, 507–16.

D. K. Foley and M. Sidrauski, *Monetary and Fiscal Policy in a Growing Economy*, New York 1971.

R. Hall and D. Jorgenson, "Tax Policy and Investment Behavior," *Amer. Econ. Rev.*, June 1967, *57*, 391–414.

M. J. Hamburger, "Interest Rates and the Demand for Consumer Durables," *Amer. Econ. Rev.*, June 1967, *57*, 1131–53.

A. C. Harberger, *The Demand for Durable Goods*, Chicago 1960.

A. A. Hirsch, M. Liebenberg, and G. R. Green, "The BEA Quarterly Econometric Model," Bureau of Economic Analysis staff pap. no. 22, U.S. Dept. of Commerce, July 1973.

F. T. Juster and P. Wachtel, (1972a) "Anticipatory and Objective Models of Durable Goods Demand," *Amer. Econ. Rev.*, Sept. 1972, *62*, 564–79.

——— and ———, (1972b) "Inflation and the Consumer," *Brookings Papers*, Washington 1972, *1*, 71–122.

T.-C. Liu and E.-C. Hwa, "Structure and Applications of a Monthly Econometric Model of the U.S.," *Int. Econ. Rev.*, June 1974, *15*, 328–65.

M. D. McCarthy, *The Wharton Quarterly Econometric Forecasting Model—Mark III*, Philadelphia 1972.

F. S. Mishkin, "Monetary Policy and Liquidity: Simulation Results," unpublished paper, July 1975.

F. Modigliani, "Monetary Policy and Consumption," in *Consumer Spending and Monetary Policy: The Linkages*, Fed. Reserve Bank Boston, Monetary Conference Series, No. 5, Boston, June 1971, 9–84.

R. Stone and D. A. Rowe, "The Market Demand for Durable Goods," *Econometrica*, July 1957, *25*, 423–43.

J. Tobin, "A General Equilibrium Approach to Monetary Theory," *J. Money, Credit, Banking*, Feb. 1969, *1*, 15–29.

A. Zellner, D. S. Huang, and L. C. Chau, "Further Analysis of the Short-Run Consumption Function with Emphasis on the Role of Liquid Assets," *Econometrica*, July 1965, *33*, 571–81.

Survey Research Center, *Survey of Consumer Finances*, various years.

# Monopoly and the Rate of Extraction of Exhaustible Resources

By JOSEPH E. STIGLITZ*

In some recent discussions of the "energy crisis," the suggestion has been put forward that the oil producing countries or the oil companies have been acting collusively and have forced the price of oil to a level far higher than it would have been in competitive equilibrium. The object of this paper is to compare the rate of exploitation of an exhaustible natural resource in competitive markets with that of a profit maximizing monopolist. The basic result of my analysis is that there is a very limited scope for the monopolist to exercise his monopoly power; indeed, under the natural "first approximation" of constant elasticity demand schedules, with zero extraction costs, monopoly prices and competitive equilibrium prices will in fact be identical. In other cases there is some tendency for a monopolist to be more "conservation minded" than a competitive market would be.[1]

## I. A Two-Period Model

The basic intuition behind my result may easily be seen. First, consider a two-period problem. Assume that there are zero extraction costs. We have a fixed stock of oil to

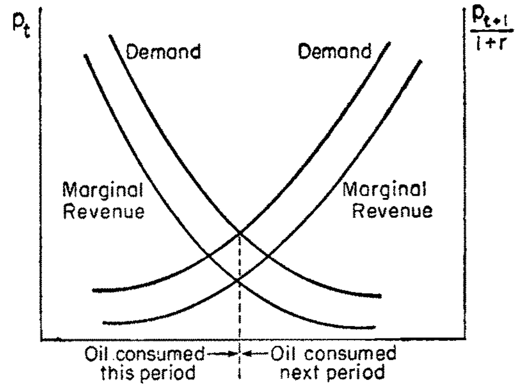



FIGURE 1

divide between two periods. That part of the stock which we do not consume the first period will be consumed the second. In Figure 1, I have plotted the demand curve this period, from the left, and the demand curve for next period, deflated by $1+r$, where $r$ is the rate of interest, from the right. In competitive equilibrium, an individual must be indifferent between selling a unit of oil today or tomorrow, so $p_t = p_{t+1}/(1+r)$. Thus, market equilibrium is the point of intersection of the two demand curves.

A monopolist, on the other hand, compares the marginal revenue he obtains this period with the marginal revenue, discounted by $1+r$, he would obtain next period by transferring a unit of sales from this period to next. In Figure 1, I have drawn the corresponding marginal revenue schedules, and the monopoly equilibrium is the intersection of the two. As I have drawn the curves, they intersect at exactly the same value of sales this period as did the price schedules, i.e., the monopoly equilibrium and the competitive equilibrium are identical. Clearly, if we have constant elasticity demand schedules, then price will be

* This work is an extension of research originally financed by the Energy Policy Project of the Ford Foundation. This work is supported in part by National Science Foundation Grant SOC74-22182 at the Institute for Mathematical Studies in the Social Sciences, Stanford University. This paper was originally written while I was a visiting fellow at St. Catherine's College, Oxford, 1973–74. Research support of the National Science Foundation and Ford Foundation is gratefully acknowledged. I am indebted to Partha Dasgupta, R. Zeckhauser, Geoffrey Heal, and an anonymous referee for discussions on the questions raised here. J. Sweeney has independently provided a similar analysis.

[1] This result is referred to in Robert Solow without a precise statement of the conditions under which it obtains. Milton Weinstein and Richard Zeckhauser establish the optimality of the competitive market's depletion of natural resources.

proportional to marginal revenue, and the two equilibria will be the same. If the elasticity of demand next period is higher than this period, the ratio of marginal revenue to price will be higher next period than this period which means that at the competitive price, discounted marginal revenue next period exceeds marginal revenue this period, so it pays to sell more next period: the monopolist is more conservationist than the competitive market. Conversely if the elasticity next period is lower than this period.

With extraction costs, the condition for competitive equilibrium is that rents, i.e., price minus extraction costs, $c$, rise at the rate of interest, i.e.,

$$p_t - c = \frac{p_{t+1} - c}{1 + r}$$

while the corresponding monopoly condition is that net marginal revenues rise at the rate of interest,

$$MR_t - c = \frac{MR_{t+1} - c}{1 + r}$$

Clearly, with constant elasticity demand schedules, since marginal revenues are a fraction of price, net discounted marginal revenue next period is greater than that for this period, when discounted rents (price minus extraction costs) next period equal that of this period. Again it pays to contract sales this period and expand them next period; with positive extraction costs and constant elasticity demand schedules, a monopolist is more conservation minded than is socially optimal.

As I shall show below, these basic results admit of considerable generalization. The basic argument is a simple one: the monopolist, like the competitor, eventually will exhaust all of the natural resource. It is not like a conventional commodity, where the total amount that will eventually be sold is smaller for a monopolist than for a competitor. Here, the only question is whether a monopolist can rearrange the patterns of sales over time to increase the present discounted value of his profits. My analysis suggests that his power to do this may be severely limited.

In a multiperiod model with zero extraction costs, competitive equilibrium will entail price rising at the rate of interest, while monopoly will require the marginal revenue to rise at the rate of interest. But if there is a constant elasticity of demand, price is proportional to marginal revenue, so price also is rising at the rate of interest. Since equilibrium entails exhaustion of the stock of resources as time approaches infinity, the competitive market equilibrium and the monopoly are described by exactly the same set of equations: the two equilibria are identical. This will be shown more formally in the next section.

## II. The Basic Model: Zero Extraction Costs, Infinite Time Horizon

Let the demand function for a quantity, $q$, of the natural resource, be of the form,[2]

$$p = f(t)q^{\alpha-1}, \qquad 1 > \alpha > 0 \tag{1}$$

where $1/(1-\alpha)$ is the elasticity of demand. The monopolist wishes to

$$\max \int_0^\infty p(t)q(t)e^{-rt}dt \tag{2}$$

subject, of course, to the constraint on the total stock of the resource, $S_0$,

$$\int_0^\infty q(t)dt \leq S_0 \tag{3}$$

Substituting (1) into (2), and introducing $\lambda$ as the Lagrange multiplier on the constraint (3), our maximization problem may be reformulated as

$$\max \int_0^\infty [f(t)q^\alpha e^{-rt} - \lambda q]dt \tag{4}$$

implying that we set $q(t)$ so that

$$\alpha e^{-rt}fq^{\alpha-1} - \lambda = 0 \tag{5}$$

which, upon substituting (1) and differentiating logarithmically, yields

[2] Obviously, if $\alpha<0$, one can obtain larger profits by reducing $q$. Some have suggested that the demand for oil in the very short run has less than unitary elasticity, but whether it is optimal for the monopolist to raise its price in these circumstances depends on the long-run demand elasticity as well. See Edmund Phelps and Sidney Winter.

$$\frac{\dot{p}}{p} = r \tag{6}$$

which is identical to the familiar condition for the time path of the price of a natural resource in a competitive market: the price must rise at the rate of interest. Again using (1), this is equivalent to

$$\frac{\dot{q}}{q} = \frac{r - \dfrac{f'}{f}}{\alpha - 1} \tag{7}$$

Thus, both the competitive market and the monopolist will satisfy the differential equation (7) and the condition (3).[3] This implies that prices at each moment of time in the competitive and monopolistic markets are identical, and hence so must be the rate of utilization of the natural resource. Because of our assumption of constant elasticity, so long as $f(t)>0$, $q>0$. Hence, if $f(t)>0$ for all time, the resource is used up only asymptotically. On the other hand, if $f(t)=0$ for $t\geq T$, then the resource is used up—in both the competitive and monopolistic markets—at exactly date T.

### III. Increasing Elasticity of Demand

There were two assumptions which were crucial to my result of the previous section—a constant elasticity which did not change over time, and zero extraction costs. In this and the next section, these two restrictions are removed.

If the elasticity of demand increases over time—as we might expect—as a result of the discovery of good substitutes for the given resource,[4] we obtain exactly the same equation for the optimal value of $p(t)$ as before (equation (5)). Differentiating (5) with respect to time, we obtain

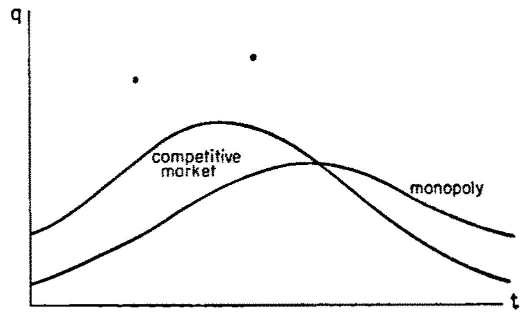



FIGURE 2. INCREASING ELASTICITY OF DEMAND

$$\frac{\dot{p}}{p} = r - \frac{\alpha'}{\alpha} \tag{8}$$

where, by assumption, $\alpha'>0$. The rate of increase of the price will be slower for the monopolist than that in the competitive market.[5] This in turn implies that if (3) is to be satisfied, the rate of utilization of the natural resource initially will be lower for the monopolist—the monopolist takes a more conservationist policy. Figure 2 compares a possible time profile of the utilization of the resource in the two markets.[6]

### IV. Extraction Costs

A similar bias for a monopolist to follow an excessively conservationist policy emerges when extraction costs are taken into account. Let the extraction cost be constant per unit of extraction but be declining with time.

[3] Equation (3) is essentially a boundary value condition.

[4] In the limiting case, where the new substitute is available with an infinite elasticity of supply the moment after discovery, if the date of discovery is known, the competitive and monopoly equilibria are identical. The case we have in mind here is where, say, a substitute is either not a perfect substitute, or does not have a perfectly elastic supply.

[5] Clearly, a more interesting case is that where the change in the elasticity of demand is an endogenous variable, say, a function of the price charged in the market. This turns out to be a far more complicated question, a special case of which is examined by Dasgupta and the author.

[6] Obviously, our formulation still is not as general as it might be, that is, within every period we assume constant elasticity demand curves. More generally, if the revenue function is of the form $R(q, t)$, then while in the monopoly market

$$\frac{\dot{q}}{q} = \frac{\left(r - \dfrac{R_{qt}}{R_q}\right)}{\dfrac{R_{qq}q}{R_q}}$$

in the competitive market

$$\frac{\dot{q}}{q} = \frac{\left(r - \dfrac{R_t}{R}\right)}{\dfrac{qR_q}{R} - 1}$$

Thus we let $g(t)$ = unit extraction cost at time t, $g' \leq 0$. The monopolist's profits are now just:[7]

$$(9) \qquad \int_0^\infty (fq^\alpha - gq)e^{-rt}dt$$

Profit maximization entails

$$(10) \qquad e^{-rt}(\alpha f q^{\alpha-1} - g) - \lambda = 0$$

which, upon differentiation and rearrangement, becomes

$$(11) \qquad \frac{\dot{p}}{p} = r(1 - \gamma_m) + \frac{\dot{g}}{g}\gamma_m$$

where

$$(12) \qquad \gamma_m = \frac{g}{\alpha p}$$

$\gamma_m$ is extraction costs divided by marginal revenue. It is clear that $\gamma_m$ must be less than unity; if extraction costs are falling rapidly, or $\gamma_m$ is large, then the market price may actually fall.

In contrast, the competitive solution requires that the individual be indifferent between extracting the oil today, receiving a net revenue (per barrel, say) of

$$p(t) - g(t)$$

or holding the oil one more period and extracting it next period, receiving a net amount of

$$\frac{p(t+1) - g(t+1)}{1+r}$$

i.e.,

$$\frac{p(t+1) - p(t)}{p(t)} = \frac{g(t+1) - g(t)}{g(t)} \cdot \frac{g(t)}{p(t)} + \frac{r(p(t) - g(t))}{p(t)}$$

or, in continuous time

$$(13) \qquad \frac{\dot{p}}{p} = r(1 - \gamma_c) + \frac{\dot{g}}{g}\gamma_c$$

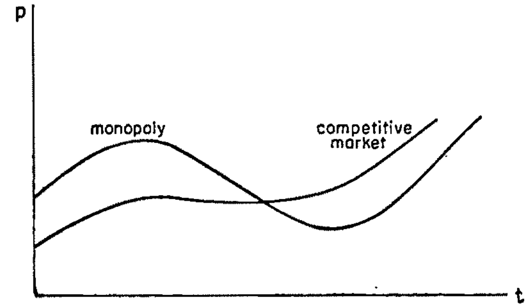




FIGURE 3. DECLINING EXTRACTION COSTS

where

$$(14) \qquad \gamma_c = \frac{g}{p}$$

Thus if at any t, $p(t)$ were the same for the competitive and monopoly markets, $\gamma_c(t) < \gamma_m(t)$ so $\dot{p}/p$ for the competitive market is greater than for the monopoly market. Thus, the "price curves" can only cross once, and the monopolist takes a more conservationist policy.[8,9] See Figure 3.

Note that if the monopolist has a lower

[7] We revert for simplicity to our assumption that $\alpha$ is constant.

[8] It is possible to show that asymptotically both the monopolist and competitor will use up the entire stock.

[9] The above formulation can be made somewhat more general by letting the extraction costs depend on the stock remaining, i.e., the costs of obtaining a flow $q$ from a field with stock $S$ at time t is $g(t)h(S, q)$. Assume we had a large number ($N$) of identical oil fields. If each owner acted competitively he would

$$\text{(a)} \qquad \text{maximize} \int_0^\infty [p(t)q(t) - g(t)h(S, q)]e^{-rt}dt$$

so, forming the Hamiltonian

$$\text{(b)} \qquad He^{rt} \equiv pq - gh(S, q) - vq$$

we obtain the result that $q$ must satisfy

$$\text{(c)} \qquad p - gh_q = v$$

while

$$\text{(d)} \qquad \dot{v} = rv + gh_S$$

The corresponding problem for the monopolist is to

$$\text{(e)} \qquad \text{maximize} \int_0^\infty [fq^\alpha N^{\alpha-1} - gh(S, q)]e^{-rt}dt$$

rate of output initially than a competitive industry, the monopolist's price must eventually be lower, if total supply is to be used up. Thus although the present generation pays higher prices for oil, subsequent generations will benefit. The monopoly equilibrium, however, is dynamically inefficient. If the monopoly were eliminated, the present generation could compensate the future generation for the higher prices, and still be better off.

## V. Speculators and Mixed Markets

In some of the cases depicted above, for example, where the elasticity of demand was falling, the price of the natural resource was rising faster than the rate of interest. This would provide an incentive for a speculator to purchase the natural resource and store it, provided storage costs were not too large. In the limit, if storage costs were zero, price would have to rise at the rate of interest, and even though the monopolist would like to be profligate with society's resources, consuming them too quickly, speculators will prevent him from doing so. Thus the monopoly and competitive equilibria are identical.

A converse argument does not hold if price is rising more slowly than the rate of interest. But if there is a mixed market, with one large holder of oil stocks and a large number of small holders, then in equilibrium the small holders will extract their oil first, with the price rising at the rate of interest; subsequently when all of their stocks are exhausted the large producer will extract, with the price rising more slowly than the rate of interest.[10]

## VI. Other Biases in the Rate of Extraction

There may, of course, be other differences between a monopoly and a competitive market. In particular, the required rate of return may be different; the monopolist for instance may have easier access to the capital market, and because of his larger size, be better able to pool risks. These suggest that the monopolist might have a lower required rate of return on capital (i.e., $r$ is smaller), which again implies a more conservationist policy for the monopolist than for the competitive market.

In any of the cases where a monopolist is conservation minded, if an industry which was previously competitive becomes cartelized, the effect will be a discontinuous jump in the price.

Whether the recent jump in the price in oil can be attributed to the factors discussed in this paper remains a moot question. It might be argued that in this case, the governments involved have less access to the capital market than do the large oil companies, so that the relevant rate of interest after cartelization was higher; on the other hand, if the oil companies had thought that there was a significant probability of nationalization, they would have pursued a policy of excessively fast extraction.

Similarly, if the rate of interest facing different firms (countries) is different, then the rate at which they would like to extract the natural resource will be different. It is clear that market equilibrium will entail the firm with the highest rate of interest extracting first (with price rising at his interest rate while he is the producer); then the next

---

so

(f) $$\alpha p - gh_q = v$$

where

(g) $$\dot{v} = rv + gh_S$$

To see clearly the difference between the two solutions, let $h = \phi(S)q$. Then for the competitive market,

$$\frac{\dot{p}}{p} = r(1-\hat{\gamma}_c) + \hat{\gamma}_c \frac{\dot{g}}{g}$$

where now $$\hat{\gamma}_c = \frac{g\phi(S)}{p}$$

while for the monopoly (letting $\hat{\gamma}_M = g\phi(S)/\alpha p$)

$$\frac{\dot{p}}{p} = r(1-\hat{\gamma}_M) + \hat{\gamma}_M \frac{\dot{g}}{g}$$

i.e., with the modification in the definition of $\gamma$, we have the same equations as before. Again, it can be shown that the monopolist pursues a more conservationist policy.

[10] This can be viewed as a Stackleberg equilibrium in which the large firm is the leader, and knows that the small firm will behave competitively; for an analysis of the Nash-Cournot equilibrium of this market, see Steve Salant.

highest, etc. If different firms face different extraction costs, the firm with the lowest extraction cost will extract first, then the next, etc. (i.e., it always pays to postpone postponable costs into the future). The competitive market equilibrium ensures that this will happen; the monopolist would behave in an identical way.

The fact that different firms with different extraction costs and apparently different rates of discount are producing simultaneously can then be attributed to: (a) marginal extraction costs are the same, even though average extraction costs are not; (b) offsetting effects of extraction costs and rates of time preference, with low extraction costs being associated with low rates of interest; (c) firms (countries) do not face a constant interest rate at which they can borrow and lend (invest); (d) risk; more particularly, differences in attitudes towards and judgments of the risks involved in postponing extraction.

Tax policy has provided further biases in the rate of extraction between the market solution and the optimal rate of extraction, but the most important provisions—the special treatment of capital gains and the depletion allowances—may not affect the relative rates of extraction of monopoly and competition. If extraction costs were zero, a constant depletion allowance would have no effect on intertemporal allocation (since price is rising at the rate of interest, the value of the depletion allowance, in present discounted terms, is independent of when the oil is depleted); hence, with constant elasticity of demand, the intertemporal resource allocation of monopoly and competition with and without the depletion allowance are all identical. With positive extraction costs, the depletion allowance encourages excessively fast depletion (since rents are rising at the rate of interest, prices are rising more slowly than the rate of interest, and hence the present discounted value of the depletion allowance is declining. Since prices with monopoly are rising more slowly than with competition, there is some presumption that the effect of the depletion allowance in accelerating extraction will be more marked in the former than in the latter. In any case, to the extent that the depletion allowance serves to offset the excessively conservative bias of monopoly, the depletion allowance may actually serve to increase social welfare. Since the return to holding a stock of a natural resource is a capital gain, and capital gains are taxed at a preferential rate, the equilibrium rate of increase of prices[11] is $r(1-t_p)/(1-t_{cg})$ where $t_p$ and $t_{cg}$ are the personal and capital gains tax rate; hence the preferential treatment of capital gains leads to excessive conservation.

A further potential source of bias is related to uncertainty, which we have ignored in this paper. An explicit treatment of the effect of monopoly when there is uncertainty is contained in Partha Dasgupta and the author.

Finally, we note that any analysis of the oil markets in the real world should probably entail an analysis of the behavior of oligopolistic markets. This would clearly take us beyond the scope of this paper; it is my hope, however, that the insights gained from comparing the polar cases of monopoly and competition will be of value in the study of these more realistic market situations.

[11] Assuming the rights for the oil under the land were acquired at essentially zero cost; otherwise, we have to take account of the tax reduction from the write-off of the value of oil rights upon exhaustion of the oil.

## REFERENCES

**P. Dasgupta and J. E. Stiglitz**, "Uncertainty and the Rate of Extraction Under Alternative Arrangements," Inst. Mathemat. Stud. in the Soc. Sci. (*IMSSS*), tech. rep. no. 179, Sept. 1975.

**E. Phelps and S. Winter**, "Optimal Price Policy Under Atomistic Competition," in Phelps, ed., *Microeconomic Foundations of Employment and Inflation Theory*, New York 1975.

**S. Salant**, "Nash-Cournot Equilibrium for an Exhaustible Resource Like Oil," mimeo., Fed. Res. Board, 1975.

**R. Solow**, "The Economics of Resources or the Resources of Economics," *Amer. Econ. Rev. Proc.*, May 1974, *64*, 1–14.

J. Stiglitz, "Taxation and the Oil Industry," report prepared for the Ford Foundation Energy Policy Project, 1973.

———, "The Efficiency of Market Prices in Long Run Allocations in the Oil Industry," in G. Brannon, ed., *Studies in Energy Tax Policy*, Cambridge 1975, 55–99.

J. Sweeney, "Economics of Depletable Resources: Market Forces and Intertemporal Bias," mimeo., Stanford Univ. 1974.

M. Weinstein and R. J. Zeckhauser, "The Optimal Consumption of Depletable Natural Resources," *Quart. J. Econ.*, Aug. 1975, *89*, 371–92.

# Specification Error in Macro-Econometric Models: The Influence of Policy Goals—Comment

By STEPHEN M. GOLDFELD*

In recent years, there has been a rapidly growing literature on various technical aspects of macro-economic policy. This has included both theoretical work on the optimal settings for policy instruments and empirical work which, by the estimation of "reaction functions," is aimed at uncovering whether actual historical policies have been systematically related to posited macro-economic targets or goals.[1] In a related strand of this literature, a number of writers have examined the consequences of systematic policy for the estimation of macro-econometric models. An article in this vein by James Crotty, which recently appeared in this *Review*, suggests that "the purposeful nature of economic policy," will typically lead model builders astray. More specifically, Crotty purports to demonstrate that systematic policy making will generate economic data which will lead model builders to obtain rather biased and perhaps even useless results.

While the spirit of this result is broadly consistent with the findings of Levis Kochin, and Alan Blinder and the author, there is one disturbing feature of Crotty's work. In particular, Crotty, following Henri Theil, posits that policy makers choose instrument settings by minimizing a quadratic loss function subject to known linear constraints, i.e., we are in a certainty-equivalence world. For such a case, however, Blinder and I have shown that model builders will encounter none of the difficulties to which Crotty alludes. The purpose of this note is to examine this discrepancy. To anticipate our results, we find that Crotty's formal demonstration is flawed by a simple statistical confusion and that in this specific context, systematic policy making presents no problems for the model builder.

## I. The Basic Model

Following Crotty, we denote the vectors of instrument and goal variables by $X$ and $Y$, respectively. The desired values of these variables are given by $X^*$ and $Y^*$. The policy problem is to choose $X$ so as to minimize

$$(1)\quad W = \alpha'(X - X^*) + \beta'(Y - Y^*) + 1/2[(X - X^*)'A(X - X^*) + (Y - Y^*)'B(Y - Y^*)]$$

subject to

$$(2)\qquad Y = RX + S$$

where (2) is the reduced form of the economy and for the moment we follow Crotty in assuming $S$ to be a set of fixed parameters.

This minimization problem is solved by substituting (2) into (1) and differentiating with respect to $X$ to yield the first-order condition

$$(3)\quad \alpha + A(X - X^*) + R'\beta + R'B(RX + S - Y^*) = 0$$

It will be noted that (3) does not involve $Y$, but Crotty substitutes $Y$ for $RX+S$ to get

$$(4)\quad \alpha + A(X - X^*) = - R'\beta - R'B(Y - Y^*)$$

which is then rewritten as[2]

* Professor of Economics, Princeton University. I would like to thank my colleagues Alan Blinder, Gregory Chow, and Richard Quandt for helpful comments.

[1] See Gregory Chow for a good review of optimal policy. A good example of the reaction function approach is contained in Richard Froyen.

[2] I have followed Crotty in assuming that the inverse in (5) exists. My equation (5) corresponds to equation (6) in Crotty.

(5) $Y = \delta - (R'B)^{-1}AX$

where $\delta = -(R'B)^{-1}(\alpha + R'\beta) + Y^* + (R'B)^{-1}AX^*$

Crotty next observes that we now have two relationships, i.e., (2) and (5), which must be satisfied, and that coefficients of a regression relating $Y$ to $X$ will not be very illuminating since they will in general reflect both relationships. More particularly, he draws the analogy with the conventional supply and demand identification problem to suggest that empirical estimates obtained by regressing $Y$ on $X$ will be dominated by whichever relationship is relatively more stable. Crotty attempts to illustrate these points in the context of a simple two goal-two instrument example. Basically, what he does is to write out (5) explicitly and show that it is different from (2). Crotty then states that if we regress $Y$ on $X$ and "assume that the regression coefficients are estimates of the optimality conditions" (i.e., (5)), then we can "judge the extent of the maximum specification bias" (p. 1027). As we shall see in a moment, one is hardly free to "assume" anything of this sort, and in fact the entire "bias" problem raised by Crotty is invalid.

## II. Statistical Difficulties

As should be clear from this brief summary of his model, Crotty's main point is a statistical or econometric one. This suggests there should be some stochastic elements in the problem. They are, of course, already implicit in $S$ which is properly regarded as a random variable reflecting both uncontrolled exogenous variables and error disturbances (loosely speaking, the residuals). Following the certainty equivalence results of Theil, Crotty has replaced this random variable with its expected value, but it will be illuminating to introduce explicitly this random element. To bring out the consequences of this most clearly, we reexamine Crotty's findings in the context of the simplest possible problem—namely, one goal and one instrument.

The object is then to minimize[3]

(6) $w = (y - y^*)^2 + a(x - x^*)^2$

subject to

(7) $y = rx + \bar{s}$

where $\bar{s}$ is the expected value of $s$ and the "true" reduced form is

(8) $y = rx + s$

Carrying out the minimization yields an optimal $x$ given by

(9) $$x = \frac{ax^* + r(y^* - \bar{s})}{a + r^2}$$

Let us now consider regressing $y$ on $x$ where $x$ is given by (9) and we have a series of observations on both variables over time. For simplicity we assume that $\bar{s}$ is constant over the sample.[4] Furthermore, the natural presumption concerning the desired values $x^*$ and $y^*$ is that they vary nonrandomly over the sample although Crotty seems to suggest they could be constant over time.[5] From (9), this means that $x$ is deterministic and therefore it is certainly uncorrelated with the random variable $s$ in (8). Consequently, the regression $y = rx + s$ satisfies all the assumptions of least squares and consequently has all the desirable regression properties.[6] In particular, the coefficient of $x$ in such a regression should be an unbiased estimator of "the" multiplier $r$.

What this demonstrates, then, is that reduced form (or, for that matter, structural) estimation in the face of a Theil-type policy maker is perfectly straightforward. In particular, since $x$ is uncorrelated with $s$ there is no question of simultaneous equations bias. But what of Crotty's argument suggesting this possibility? As will be readily apparent, the problem with Crotty's line of reasoning is that his "extra equation" is not really

[3] For further simplification, the linear terms present in (1) are ignored in (6).

[4] This still allows for $s$ to represent both uncontrollable exogenous variables and error disturbances, but abstracts from possible trends in these exogenous variables. This is for convenience only and has nothing to do with the basic point I shall make.

[5] If both $x^*$ and $y^*$ are constant, then $x$ will be as well and one would hardly be tempted to use it as an independent variable in a regression.

[6] If $x^*$ and $y^*$ are random, then all that is required is that they be independent of $s$.

extra at all.

To see this, recall that the equation in question, (5), was obtained by replacing $RX+S$ in (3) with $Y$. There are two difficulties with this procedure. First, it is not quite correct since, as is clear from (9) what really appears in (3) is $RX+\bar{S}$ which is not $Y$. One can, of course, make the proper substitution. In the present simple case (3) takes the form

$$(10) \quad a(x - x^*) + r(rx + \bar{s} - y^*) = 0$$

Using the fact that $rx=y-s$, (10) can be rewritten as

$$(11) \quad y = y^* - \frac{a}{r}(x - x^*) + (s - \bar{s})$$

which is the corrected analog of (5).

The more important difficulty concerns the proper interpretation of (11) (or (5)). While (11) is a perfectly valid equation, the basic point is that it is a contrived and highly artificial one which is not independent of the reduced form. The easiest way to see this is to solve (9) for either $x^*$ or $y^*$ and substitute back into (11). In either case this will yield $y=rx+s$, demonstrating that (11) need not concern us.

Put another way, there is only one relationship between $y$ and $x$ which is present and that is the reduced form $y=rx+s$. Equation (11) is simply a related equation which involves two *additional* variables, $x^*$ and $y^*$.

One might, of course, estimate (11) in its own right to yield an estimate of $a/r$. This would require data on $x^*$ and $y^*$ which is generally not available. If it were, however, one could do better still by estimating (9) which must fit *perfectly*. Clearly, the ratio of the coefficients of $x^*$ to $y^*$ would give an exact estimate of $a/r$. Consequently, (11) is even superfluous from this point of view.[7]

In short, as far as the model builder is concerned, the "extra" equation is a red herring, pure and simple.

## III. Summary

This note has shown that in a Theil-type world, regressing endogenous target variables on "controlled" instrument variables will lead to unbiased estimates of the reduced form. The kinds of difficulties alluded to by Crotty, based on vague appeals to simultaneous equations problems, simply do not arise. There is nothing the policy maker does which can produce the statistical problems (correlations between instruments and disturbances) which would be needed to justify Crotty's claims. This is not to say that these problems could not arise in other circumstances, but this has been treated rather fully in the article by Blinder and the author.

[7] Equation (11) could be relevant if one is interested in the nature of the policy maker's preferences, embodied here by the value of $a$ in (6). If $x^*$ and $y^*$ were not known precisely, an exact estimate of $a/r$ would not be available. One could then fit (9) or (11), the choice depending on the nature of the errors-in-variables assumptions, to get an estimate of $a/r$ and use the reduced form estimate of $r$ to obtain an estimate of $a$. This is, in fact, the procedure followed by Anne Friedlaender.

## REFERENCES

A. Blinder and S. Goldfeld, "Some Implications of Endogenous Stabilization Policy," *Brookings Papers*, Washington 1972, *3*, 585–640.

G. C. Chow, "Problems of Economic Policy from the Viewpoint of Optimal Control," *Amer. Econ. Rev.*, Dec. 1973, *63*, 825–37.

J. R. Crotty, "Specification Error in Macro-Econometric Models: The Influence of Policy Goals," *Amer. Econ. Rev.*, Dec. 1973, *63*, 1025–30.

A. F. Friedlaender, "Macro Policy Goals in the Postwar Period: A Study in Revealed Preferences," *Quart. J. Econ.*, Feb. 1973, *87*, 25–43.

R. T. Froyen, "A Test of the Endogeneity of Monetary Policy," *J. Econometrics*, July 1974, *2*, 175–88.

L. Kochin, "Judging Stabilization Policies," unpublished doctoral dissertation, Univ. Chicago 1973.

# Specification Error in Macro-Econometric Models: The Influence of Policy Goals—Reply

By JAMES R. CROTTY*

Stephen Goldfeld's conclusions that the analysis of my article is "flawed by a simple statistical confusion" and that "the extra equation is a red herring pure and simple" are based on a set of assumptions which are virtually the opposite of the ones I worked with, and depend on an excessively restrictive interpretation of what is implied by the use of the Theil certainty equivalence framework. Under the assumptions stated in my article, my conclusions remain valid. It is Goldfeld's new assumptions that create the "errors" he then discovers in my paper.

The policy problem that I modelled was one in which the state had relatively firm policy targets and adjusted its policy instruments in response to shifts in the expected values of uncontrolled variables. In order to emphasize the potential problem, I assumed that $X^*$ and $Y^*$ were constant, so that $X$ was chosen to minimize equation (1), the quadratic loss function, as $\bar{S}$, the expected value of $S$, shifted from period to period.[1] As I stated in the article, "$S$ clearly changes every period" (p. 1027). But I used the symbol $S$ to denote the policy maker's certainty-equivalent forecast of $S$, i.e., Goldfeld's $\bar{S}$. My policy examples all contained the same assumptions. "Instrument movements would . . . counterbalance other forces tending to move the goal variables away from their desired values" (p. 1029) represents a typical sample from my discussion. Following Theil, I assumed that the policy maker behaved as if $\bar{S}$, the expected value of his subjective probability distribution (or forecast distribution) of $S$, was known with certainty, but *not* that the forecast distribution or $\bar{S}$ never shifted or changed over time.[2]

Goldfeld reverses these assumptions to describe and analyze a fickle policy maker in a world of constant expected values of uncontrolled variables, the very opposite of my assumptions and a relatively uninteresting case at that. He argues that the "natural presumption" is the $X^*$ and $Y^*$ vary nonrandomly over the sample and, "for simplicity," that $\bar{S}$ is constant over the sample. Only for the case where $\bar{S}$ is constant is Goldfeld's conclusion in footnote 5 that the constancy of $X^*$ and $Y^*$ implies the constancy of the policy instruments $X$ correct. In footnote 4 Goldfeld suggests that the assumption that $\bar{S}$ is constant is "for convenience only," but with respect to the problem addressed in my article, this is not so.

To see this, define $\epsilon_t = s_t - \bar{s}_t$.[3] Here $\epsilon$ is a random variable with an expected value of zero and constant variance, and the "true" reduced form is $y = rx + \bar{s} + \epsilon$. Under Gold-

* Assistant professor of economics, University of Massachusetts at Amherst. I am grateful to Raford Boddy and James Kindahl for helpful comments. Any errors, of course, are my responsibility.

[1] Goldfeld's notation and equation numbering will be used throughout, except where specifically noted.

[2] Rereading the original article, the only place I can find where Goldfeld could conceivably have been misled concerning my treatment of $\bar{S}$ was on page 1026 where I referred to $\bar{S}$ as a vector of "fixed elements." Here I followed Theil in viewing the noninstrument variables as uncontrolled and treated as if their values were known with certainty over the fixed planning horizon for the purpose of making the *first period's decisions*. In the article I had in mind a one-period, moving, fixed-horizon model and it is probably easiest to see the general problem under this simplifying assumption. Even if the planning horizon were assumed to be extremely long, the estimates of $\bar{S}$ elements for future periods would change as new information became available to the policy maker each period and in any case the current period values of the uncontrolled variables would not be expected to remain constant or unchanging from period to period. Furthermore, the rest of the article makes my treatment of $\bar{S}$ so clear that I find it difficult to understand Goldfeld's mix-up.

[3] Goldfeld switches from capital to small letters in going from the general statement of the problem to his one goal, one instrument example.

feld's assumptions, $x$ cannot be correlated with $\bar{s}$ because $\bar{s}$ is a constant, and $x$ is uncorrelated with $\epsilon$ by definition. The problem I addressed has thus been defined out of existence. If we return to my simplifying assumptions that $x^*$ and $y^*$ are constants but that $\bar{s}$ shifts from period to period, Goldfeld's conclusions are untenable. Rewriting his equation (9) as:

$$(9') \quad x = \left(\frac{a}{a+r^2}\right)x^* + \left(\frac{r}{a+r^2}\right)y^* - \left(\frac{r}{a+r^2}\right)\bar{s}$$

it is clear that $x$ is a linear function of $\bar{s}$ and therefore, contrary to Goldfeld's assertion, it is highly correlated with $s$ because $s=\bar{s}+\epsilon$.

Rewrite his equation (11) as:

$$(11') \quad y = \left(y^* + \frac{a}{r}x^*\right) - \frac{a}{r}x + \epsilon$$

In (11′), $x$ is uncorrelated with $\epsilon$. Therefore, a regression of $y$ on $x$ (the explicit case considered in my article and the one which typically arises in econometric studies designed to estimate policy instrument impact multipliers), will yield coefficients which are functions of both the reduced form impact multipliers, represented here by $r$, and the preference function parameters represented by $a$. Under the assumptions used in my article to show the extent of the maximum specification error, the regression coefficients will be unbiased estimates of the optimality condition, i.e., of $a/r$ in equation (11′). Therefore, almost all studies to date do have a potentially serious identification problem caused by the (omitted) influence of the preference function. In other words, if we adhere to the assumptions of my article, my conclusions are valid as originally stated.[4]

The specific estimation problem analyzed here is present to some degree as long as any controlled or uncontrolled variables are omitted from the reduced form regression and could be quite severe in Andersen-Jordan type instrument-goal regressions. Moreover, the analysis presented provides insight into some econometric difficulties involved in the estimation of the full reduced form inclusive of the uncontrolled variables—a problem not specifically addressed in my article. Suppose one regressed a goal variable against all relevant instrument and uncontrolled variables. As can be seen by inspection of equation (9′), if $X^*$ and $Y^*$ were constant over the sample, the $X$ variables would be linear combinations of the $\bar{S}$ variables and the data matrix might be very nearly singular. If $X^*$ and $Y^*$ varied over the sample, but varied significantly less than did $\bar{S}$, all the $X$ variables would be highly correlated with the $\bar{S}$ variables and with each other because of the preference function. The regression would still be riddled with multicollinearity and it might be quite difficult to disentangle the relative influences of the various independent variables.

Thus it seems reasonable to conclude that the econometric problem discussed in my article was not a "red herring" and that Goldfeld himself created the "errors" he found in my work.

[4] Partly because model builders have been unconcerned with this issue, we have as yet little or no hard econometric information concerning policy preference functions. In footnote 7 Goldfeld suggests that a satisfactory procedure for the estimation of preference function parameters has been presented by Anne Friedlaender. Her work in this area has been exceptionally creative but she is justifiably cautious concerning the robustness of her estimates and the reliability of her estimation procedures. See Boddy and the author for a discussion of some of the problems involved in the Friedlaender procedure, and the author for an alternative procedure which overcomes some limitations of her approach but shares many of the same difficulties.

## REFERENCES


**R. Boddy and J. Crotty**, "Macro-Policy Goals in the Postwar Period: Comment," *Quart. J. Econ.*, Nov. 1975, *89*, 673–76.

**J. Crotty**, (1973a) "The Objective Function of the State: A Study of the Political Economy of Monetary and Fiscal Policy," unpublished doctoral dissertation, Carnegie-Mellon Univ. 1973.

———, (1973b) "Specification Error in Macro-Econometric Models: The Influence of Policy Goals," *Amer. Econ. Rev.*, Dec. 1973, *63*, 825–37.

A. F. **Friedlaender**, "Macro-Policy Goals in the Postwar Period: A Study in Revealed Preferences," *Quart. J. Econ.*, Feb. 1973, *87*, 25–43.

S. M. **Goldfeld**, "Specification Error in Macro-Econometric Models: The Influence of Policy Goals—Comment," *Amer. Econ. Rev.*, Sept. 1976, *66*, 662–64.

# Commodity Trade and Factor Mobility: Comment

*By* TERUTOMO OZAWA*

In a recent paper in this *Review*, Melvyn Krauss presented an interesting case of factor mobility under the assumption that two countries have an identical factor endowment but differ from each other in tastes. A pattern of factor movements required under this set of assumptions is shown to be determinate in contrast to the basic indeterminacy involved in a situation that assumes differences in factor endowments but identical tastes in two countries.

To prove his point that there is only one unique pattern of factor mobility, Krauss starts by drawing from the origin of a two-factor endowment (capital and labor) diagram two rays representing "the factor endowment ratios in the two countries that correspond to the production vectors . . . required for the mirror image solution" (p. 801) in such a manner that each ray symmetrically deviates from the initial common factor endowment ratio line from the origin. He then proves with the help of a parallelogram the exact amounts of labor and capital to be transferred between the two countries.

Intuitively we know that such factor endowment ratio rays must exist. Yet we are not in a position to tell the exact locations of such rays unless we first know the exact amounts of labor and capital to be moved between the two countries. Although Krauss's arbitrary way of drawing such rays does not invalidate his proof, it seems more logical as a geometrical proof that we directly find the exact amounts of two factors to be transferred in a manner shown below, since only after we have done so, can the rays used by Krauss be derived.

The initial common factor endowment box is shown by $O_yCO_xL$ in Figure 1. The equilibrium production point achieved under free trade is indicated by point $P$. Let us subtract the amount of labor-intensive good $Y$ which the home country would export

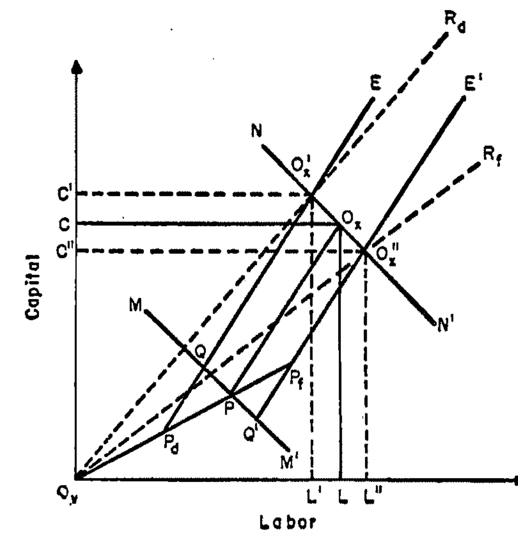



FIGURE 1

under free trade, namely $PP_d$. Draw a line $P_dE$ parallel to $PO_x$ and indicate the amount of capital-intensive good $X$ which the home country would import under free trade, namely $P_dQ$. Then draw a line $MM'$ connecting $Q$ and $P$. The home country's "commodity trade triangle" is $PP_dQ$. We can similarly construct the foreign country's commodity trade triangle, $PP_fQ'$, and the line $P_fE'$.

Next draw a line $NN'$ through point $O_x$ parallel to the line $MM'$, and identify its intersections with $P_dE$ and $P_fE'$ as points $O_x'$ and $O_x''$, respectively. Since $PQ=PQ'$, $O_xO_x'$ must equal $O_xO_x''$. We have thus found that the home country must "import" $CC'$ of capital and "export" $L'L$ of labor, while the foreign country imports $LL''$ of labor and exports $CC''$ of capital, and that any other pattern of factor movements is impossible. We are now in a position to draw the rays which are used by Krauss, as shown by dotted lines $O_yR_d$ and $O_yR_f$, by connecting $O_y$ and $O_x'$ for the home country and $O_y$ and $O_x''$ for the foreign country, respectively.

* Professor of economics, Colorado State University.

# Commodity Trade and Factor Mobility: Reply

By Melvyn B. Krauss*

Terutomo Ozawa's demonstration of the relationship between the commodity trade triangle, the factor trade triangle, and the equilibrium factor endowment ratio rays in two countries that initially trade with one another because of differences in tastes, is both logically correct and geometrically interesting. But while I admire his elegant proof, I must disagree with his assertion that "we are not in a position to tell the exact locations of [the factor endowment ratio] rays unless we first know the exact amounts of labor and capital to be moved between the two countries" (p. 668). This assertion not only runs counter to the general mathematical fact that the solution of a general equilibrium system requires that all unknowns be simultaneously determined, but as a specific proposition is demonstrably false as well.

It is proved below that given the initial equilibrium commodity trade triangles, the equilibrium factor endowment ratio rays in the two countries can be determined without *explicitly* considering the exact amounts of labor and capital to be moved between the two countries. In Figure 1 (a simplified version of my earlier Figure 2), the initial common production vector is $OP$ with $OZ_f$ the consumption vector in the foreign country and $OZ_d$ the consumption vector in the home country. The respective equilibrium triangles thus are $C_fFP$ and $C_dDP$, and the equilibrium commodity price ratio given by the slope of $C_fPC_d$.

By the Stolper-Samuelson relation the factor price ratio is known once the commodity price ratio is known; and given factor prices, the optimal factor-intensity ratios $OR_x$ and $OR_y$ also are known. In Figure 2, point $E$ represents the common factor endowment

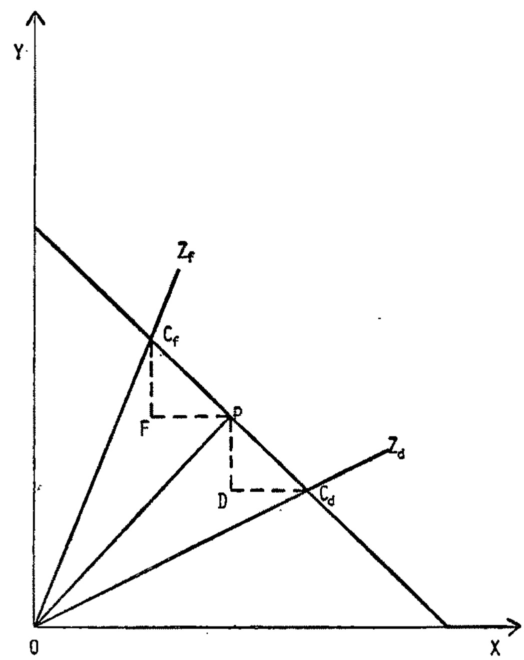



Figure 1

point in the two countries. By vector addition, the common factor allocation points $P_x$ and $P_y$ are determined. The ratio of $X$ production to $Y$ production is reflected by the slope of the straight line connecting points $P_x$ and $P_y$. This ratio is equal to the slope of $OP$ in Figure 1.

For factor mobility to cut off all goods traded in this case, the production vectors in each country must coincide with the respective consumption vectors—that is, the foreign production vector must coincide with $OZ_f$ and the domestic production vector must coincide with $OZ_d$. This implies that at constant international prices factor mobility must increase the ratio of $X$ production to $Y$ production in the domestic country and reduce that ratio in the foreign country. The precise ratios of $X$ production to $Y$ production in both countries that are required for

* Visiting research professor, Institute for International Economic Studies, Stockholm. The paper was written while I was visiting associate professor of economics, Stanford University.

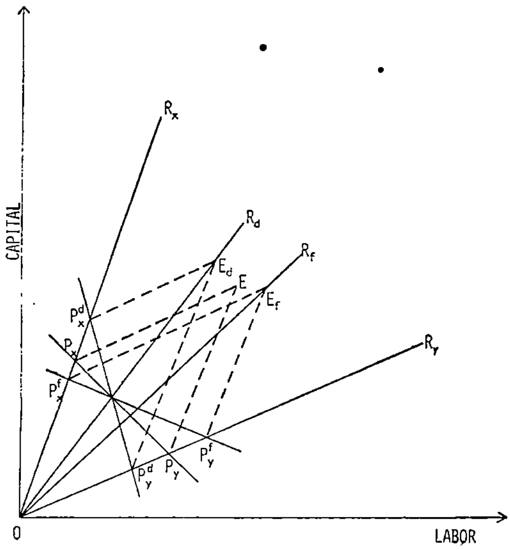



FIGURE 2

equilibrium with factor mobility are known, since the consumption vectors in both countries are known. The ratio corresponding to $OZ_d$ is assumed to be reflected by the slope of $P^d_xP^d_y$ in Figure 2, since the slope of $P^d_xP^d_y$ indicates a higher ratio of $X$ production to $Y$ production than the slope of $P_xP_y$. Similarly, the ratio corresponding to $OZ_f$ is assumed to be reflected by the slope of $P^f_xP^f_y$, which indicates a higher ratio of $Y$ production to $X$ production than the slope of $P_xP_y$.

Given the production ratio lines $P^d_xP^d_y$ and $P^f_xP^f_y$, and the factor-intensity ratios $OR_x$ and $OR_y$, the factor endowment point $E_d$ consistent with $P^d_x$ and $P^d_y$ and the factor endowment point $E_f$ consistent with $P^f_x$ and $P^f_y$ are determined. Points $E_d$ and $E_f$ are not necessarily the equilibrium factor endowment points, since $P^d_x$, $P^d_y$, $P^f_x$, and $P^f_y$ are not necessarily the equilibrium factor allocation points (the slope of $P^d_xP^d_y$ is the equilibrium production ratio but $P^d_x$ and $P^d_y$ need not represent the equilibrium absolute outputs of $X$ and $Y$ in the home country). But both $E_d$ and $E_f$ must lie on the equilibrium factor endowment rays in their respective countries. Hence the equilibrium factor ratios $OR_d$ and $OR_f$ are determined without explicit consideration of the exact amounts of labor and capital to be moved between the two countries. Rather than being arbitrarily drawn, as Ozawa claims, the rays $OR_d$ and $OR_f$ in my earlier Figure 3 are determined from the commodity trade triangle under the assumptions of the Heckscher-Ohlin-Samuelson model.[1] Except for certain expositional differences, Ozawa's and my own proof are substantially the same.

[1] Given $OR_d$, $OR_f$, and $E$, the equilibrium pattern of factor flows can be determined by my earlier Figure 3. Once the equilibrium factor endowments are determined for each country, the equilibrium factor allocation points also are determined by the technique demonstrated in Figure 2.

## REFERENCES

**M. B. Krauss**, "Commodity Trade and Factor Mobility," *Amer. Econ. Rev.*, Sept. 1974, *64*, 797–801.

**T. Ozawa**, "Commodity Trade and Factor Mobility: Comment," *Amer. Econ. Rev.*, Sept. 1976, *66*, 668.

# Can Trade Widen the Difference Between Factor Rewards? Another Look at the More-Goods-Than-Factors Case

By DOUGLAS B. STEWART*

Since publication of the seminal articles of Jan Tinbergen, Paul Samuelson (1949, 1953), and James Meade it has become increasingly clear that the traditional international trade propositions associated with a two-good two-factor world generally do not hold unchanged in a three-good two-factor world. Moreover, the modifications required to extend those propositions to the three-good world have proven so far to be neither very intuitive nor straightforward.[1] These difficulties have been evidenced in the trade theory literature by continuing and nontrivial disagreements on exactly how the three- and two-good worlds differ. In this context the important and clashing papers by A. H. Land and Harry Johnson (1967) stimulated further work and debate on the conditions for factor-price equalization. The latest word here seems to be the work of Jaroslav Vanek and Trent Bertrand.

In addition to factor-price equalization, a second major question was debated in the Land-Johnson exchange. Land argued by means of an example that free trade in a three-good two-factor world could cause factor-price ratios to diverge from their autarkic values. Johnson (1967) argued that Land's example was subject to certain inconsistencies and, therefore, factor-price ratios necessarily converge with trade. Strangely, this question has received no further analytical attention, although the correctness of Johnson's argument was asserted by both parties in a later exchange between Bertrand and Johnson (1970). The purpose of this note is to show that Johnson's argument on the possibility of diverging factor rewards is incorrect.

To set the stage, consider two countries both of which have endowments of capital and labor which are to be efficiently employed in the production of some mix of goods $X$, $Y$, and $Z$. Both countries face the same constant-returns-to-scale production technology. We assume country 1's endowment is relatively labor abundant and $Z$ is always the labor-intensive good while $X$ is always capital intensive. Finally, we assume positive demand for each good in each country at all relevant prices.

The behavior of a number of variables relevant under the above assumptions can be graphically analyzed. Following Land's Figure 6 (p. 141), Figure 1 here shows both countries in equilibrium in autarky and with free trade. The subscripts $iT$ and $i$ refer to country $i$ with and without trade, respectively. Given the equilibrium commodity prices in autarky and taking $X$ as numeraire, the $Y_1$ and $Z_1$ isoquants represent the quantities of $Y$ and $Z$ equal in value to one unit of $X$ in country 1. By constructing the convex envelope to the $X$, $Y_1$, and $Z_1$ isoquants we obtain the unit iso-revenue curve $xfez_1$. The equilibrium wage-rental ratio is then given by the slope of the iso-revenue curve where it is intersected by the ray $k_1$ representing country 1's factor endowment ratio. Thus $W_1$ is the unit cost line and its points of tangency with the $X$, $Y_1$, and $Z_1$ isoquants determine the optimal factor intensities in each sector; for example, the ray $of$ indicates the optimal capital-labor ratio for the production of $X$ in country 1.

The situation in country 2 is depicted similarly. Note that the equilibrium autarkic prices must be such that all three

* San Diego State University. I wish to thank Chulsoon Khang, Murray Kemp, Harry Johnson, and Jagdish Bhagwati for their helpful suggestions and encouragement on an earlier draft. I have also benefited from the comments of an anonymous referee and the managing editor. All errors here are my own.

[1] See John Chipman's survey of modern trade theory which nicely gives separate treatment to the unequal goods and factors cases in each topic area. Also see the appendices to Murray Kemp's first and third chapters which carefully generalize the Stolper-Samuelson, Samuelson-Rybczynski, and Factor-Price Equalization theorems to any number of goods and factors.

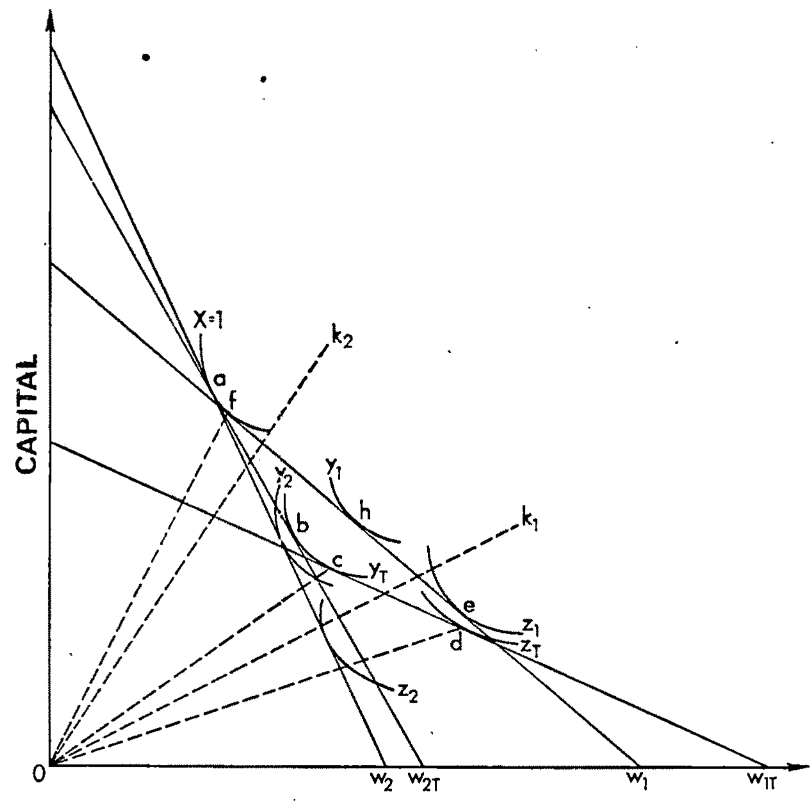



FIGURE 1

unit-value isoquants are tangent to the isocost line since all goods must be produced. Note also that the lower the price, the greater the unit-value quantity. Thus Figure 1 illustrates the case where both $Y$ and $Z$ are more expensive (relative to $X$) in country 2 than in country 1 in autarky. It follows, then, that with free trade the equilibrium terms of trade must be such that the unit value isoquants $Y_T$ and $Z_T$ lie between the pretrade unit value isoquants of the two countries. In the case illustrated in Figure 1 both countries face the unit iso-revenue curve $xabcdz_T$ in trade equilibrium. Given their respective factor endowments we see country 1 will produce only $Y$ and $Z$, and will have the wage-rental ratio given by the slope of $W_{1T}$, while country 2 will only produce $X$ and $Y$ with the wage-rental ratio given by the slope of $W_{2T}$. With trade the wage-rental ratio has fallen in each country and it has fallen more in country 1, the country with the lower pretrade wage-rental ratio. Therefore, factor-price ratios have diverged with trade, and Figure 1 illustrates the possibility as argued by Land.[2]

Johnson (1967) claims that with free trade the factor-price ratios must not only converge but one must fall and one must rise, and, thus Land's example is not a legitimate possibility. He argues as follows. Suppose the wage-rental ratio in country 2 falls with trade and in trade equilibrium it produces only the two relatively capital-intensive goods and does not produce $Z$. Now if the wage-rental ratio also falls in country 1,

[2] Land's Figure 6 (p. 141) does not exactly match her verbal argument. In her diagram the higher wage-rental ratio has fallen more than the lower one with trade. Thus they have converged rather than diverged as was intended.

which is the labor abundant country, it can cease production of $X$ and specialize in the two labor-intensive goods only if its production of $Z$ declines from the pretrade equilibrium. But this is impossible, claims Johnson, since it implies that after trade, 1) world production of $Z$ is less than before trade, and 2) country 1 is producing less of the good intensive in its relatively abundant factor (Johnson, 1967, p. 286 and p. 287, fn. 1).

Johnson's analysis of the changes brought about by trade is quite correct, yet he has not proven that any contradiction arises from them. From statements in his 1967 article as well as in his 1970 exchange with Bertrand it is clear that he sees some inconsistency between general demand conditions and the production mix determined by the supply side for the equilibrium terms of trade in Land's example. However, the assumed regular demand conditions cannot tell us what has happened to world demand for any of the commodities since no one commodity becomes either cheaper or more expensive relative to both remaining commodities as long as the terms of trade are between the pretrade prices. In particular, for the case at hand, the price of $Z$ in country 1 rises with trade. Therefore, declining world demand for $Z$ is possible, and, contrary to Johnson's claim, declining production of $Z$ raises no contradiction.[3]

Notwithstanding the above point, there are two potential inconsistencies in examples like Land's, and although her example avoids both it can be instructive to consider them briefly. First, there are some terms of trade between the pretrade prices at which one good would not be produced in either country regardless of their factor endowments. This would be inconsistent with the assumption of positive demand for each good and, hence, such terms of trade could not be equilibrium values. For example, in Figure 1 if trade caused the price of $Z$ to fall only slightly in country 2 and the price of $Y$ to rise only slightly in country 1, the unit value $Y$-isoquant would lie above the unit isorevenue curve and production of $Y$ would not be viable in either country. Second, the assumed factor endowments may make some potential equilibrium terms of trade inconsistent with world production of all three goods. For the pattern of production shown in Figure 1 to emerge it is required that the endowment rays $k_1$ and $k_2$ lie within the respective cones *cod* and *aob*. If, for example, country 1 had a less labor-abundant endowment so that the ray $k_1$ passed to the left of tangency point $h$, a rising rather than a falling wage-rental ratio would be required to induce country 1 to specialize in $Y$ and $Z$.[4]

Returning to Johnson's second point against Land's example, he claims it is impossible for country 1 to produce less of $Z$ with trade, as that result "contradicts the assumption that country 1 is led by free trade to expand the production of the good in which its factor endowment gives it a comparative advantage" (p. 287, fn. 1). But Johnson's "assumption" here is not an axiom of accepted trade theory. In the two-good two-factor world it is a readily established *result* since with free trade the price of the good in which a country has a comparative advantage must rise relative to the price of the one other good in that country and, hence, more of the good will be produced in that country. However, in the three-good two-factor world of Land's example the price of $Z$ in country 1 rises relative to the price of $X$, but falls relative to the price of $Y$. Thus, increased production of $Z$ does not necessarily follow.

In conclusion, Land's example is not subject to any inconsistencies as claimed by Johnson, and thus the divergence of factor

[3] Indeed, we could construct the demand conditions with $Z$ so strongly inferior (but non-Giffen) as to insure declining demand. However, I do not claim $Z$ inferiority is necessary for this outcome. Lower demand for $Z$ could result from various demand features such as: $Z$ highly price elastic in country 1; $Z$ highly complementary with $Y$ in country 1; $Z$ highly price inelastic in country 2; $Z$ strongly complementary with $X$ in country 2.

[4] Bertrand has used this example in correcting Johnson's (1967) treatment of factor-price equalization. In the same note he accepts Johnson's (1967) argument against diverging factor prices. And it seems at least a possibility that he was led to that acceptance by failing to note this crucial difference between his example and Land's.

rewards remains a possibility in a three-good, two-factor, two-country world. Moreover, the pattern of production and trade seems very reasonable. Free trade allows country 1 to concentrate its labor-intensive endowment in production of the two most labor-intensive goods. However, in doing so it must adjust the relative production levels of these two goods in order to maintain full factor-employment. This adjustment is such as to reduce the production of $Z$ and increase production of the good of intermediate factor intensity. Note also that the trade pattern agrees with the Heckscher-Ohlin result as country 1 must export $Z$ and country 2 must export $X$ in trade equilibrium.[5]

[5] Of course the standard Heckscher-Ohlin theorem does not hold in this model. For consideration of other possible trade patterns, see James Melvin (1968, 1971, 1972) and the author (1971, 1972).

## REFERENCES


**T. J. Bertrand,** "On Factor Price Equalization When Commodities Outnumber Factors: A Note," *Economica*, Feb. 1970, *37*, 86–88.

**J. S. Chipman,** "A Survey of the Theory of International Trade: Part 3, The Modern Theory," *Econometrica*, Jan. 1966, *34*, 18–76.

**H. G. Johnson,** "On Factor Price Equalization When Commodities Outnumber Factors: A Comment," *Economica*, Feb. 1970, *37*, 89–90.

———, "The Possibility of Factor-Price Equalization When Commodities Outnumber Factors," *Economica*, Aug. 1967, *34*, 282–88.

**M. C. Kemp,** *The Pure Theory of International Trade and Investment*, Englewood Cliffs 1969.

**A. H. Land,** "Factor Endowments and Factor Prices," *Economica*, May 1959, *26*, 137–42.

**J. E. Meade,** "The Equalization of Factor Prices; The Two-Country Two-Factor Three-Product Case," *Metroeconomica*, Dec. 1950, *2*, 129–33.

**J. R. Melvin,** "Production and Trade With Two Factors and Three Goods," *Amer. Econ. Rev.*, Dec. 1968, *58*, 1249–68.

———, "Production Indeterminacy With Three Goods and Two Factors: Reply," *Amer. Econ. Rev.*, Mar. 1971, *61*, 245–46.

———, "Production Indeterminacy With Three Goods and Two Factors: The Last Word?," *Amer. Econ. Rev.*, Sept. 1972, *62*, 723.

**P. A. Samuelson,** "International Factor Price Equalization Once Again," *Econ. J.*, June 1949, *59*, 181–97.

———, "Prices of Factors and Goods in General Equilibrium," *Rev. Econ. Stud.*, Jan. 1953, *21*, 1–20.

**D. B. Stewart,** "Production Indeterminacy With Three Goods and Two Factors: A Comment on the Pattern of Trade," *Amer. Econ. Rev.*, Mar. 1971, *61*, 241–44.

———, "Production Indeterminacy With Three Goods and Two Factors: Rejoinder," *Amer. Econ. Rev.*, Sept. 1972, *62*, 720–22.

**J. Tinbergen,** "The Equalization of Factor Prices Between Free-Trade Areas," *Metroeconomica*, Apr. 1949, *1*, 39–47.

**J. Vanek and T. J. Bertrand,** "Trade and Factor Prices in a Multi-Commodity World," in J. Bhagwati et al., eds., *Trade, Balance of Payments and Growth*, Amsterdam 1971.

# Factor Demand with Output Price Uncertainty

*By* RICHARD HARTMAN*

The effects of output price uncertainty on a competitive firm's supply and factor demands have recently been explored under the assumption that all decisions are made before the price is observed. Agnar Sandmo has shown that a risk-averse, competitive firm with a nonrandom cost function will produce a smaller output if the price is random than it would if the price were known with certainty to equal its mean. Raveendra Batra and Aman Ullah and I have extended the analysis by considering the effects of output price uncertainty on factor demands. Among other things, Batra and Ullah show that the firm will choose its inputs to minimize the cost of producing whatever level of output is chosen. This result, combined with the Sandmo result that the presence of uncertainty reduces output, implies that the effects of uncertainty on factor demands depend on what effect the decreased output due to uncertainty has on the cost minimizing levels of inputs. Except for the rare case of inferior factors, the presence of uncertainty reduces factor demands. Finally, it is clear from these analyses that if the firm is risk neutral, the uncertainty has no effect on supply and factor demands.

In this paper I relax the assumption that *all* inputs are chosen before the output price is observed. My conclusions show that the results noted above are really rather sensitive to that particular assumption. A simple two-input, one-output model of the firm is employed. One of the inputs, which I call capital, is quasi fixed in the sense that it must be chosen before the output price is observed. The other input, which I call labor, is variable since it is not chosen until the output price is observed. Clearly, this implies that the level of output is not determined until its price is known. Although labor may be a poor name for the variable input in view of the recent discussions of its "quasi-fixed" character, it seems apparent that in many situations there are inputs which can be varied on short notice.

By allowing a variable input in this sense we considerably alter the situation facing the firm. If, after observing the output price, it turns out that the firm made a "poor" choice regarding the quasi-fixed factor, it is able to partially "adjust" by choosing an appropriate level of the variable input. This ability to make adjustments for what, *ex post*, appears to be a poor decision is totally lacking if all inputs must be chosen before the uncertainty is resolved.

In Section I the basic model is presented. In Section II the analysis for a risk-neutral firm is carried out, and Section III contains an example. In Section IV the analysis is extended to a risk-averse firm. The final section contains some brief concluding comments.

## I. The Basic Model

Let us consider a price-taking firm which produces an output $Q$ using capital $K$ and labor $L$ as inputs. These variables are related by a production function

$$(1)\qquad Q = F(K, L)$$

which has positive marginal products,[1]

$$(2)\qquad F_1(K, L) > 0, \qquad F_2(K, L) > 0$$

and is strictly concave,

$$(3)\qquad F_{11}(K, L) < 0, \qquad F_{22}(K, L) < 0$$
$$F_{11}(K, L)F_{22}(K, L) - F_{12}^2(K, L) > 0$$

We take the capital input to be quasi fixed while the labor input is completely variable. By this we mean that the capital

* Assistant professor of economics, University of Washington, Seattle. I am indebted to an anonymous referee for helpful comments.

[1] Subscripts denote partial differentiation with respect to the corresponding argument.

input must be chosen before the output price $p$ is observed, but that the labor input is not chosen until after it is observed. It is assumed that the firm has some subjective probability distribution regarding $p$ before it is actually observed. Finally, we assume that the prices of the capital input $c$ and of the labor input $w$ are known with certainty.

Since the labor input is chosen after the capital input has been determined and after the output price has been observed, it is clear that labor will be chosen to maximize short-run profits

$$pF(K, L) - wL \tag{4}$$

The first-order condition for this optimization is

$$pF_2(K, L) - w = 0 \tag{5}$$

which can be solved for the optimal labor input

$$L^* = L^*(K, p, w) \tag{6}$$

Substituting $L^*$ into (4) gives the short-run profit function

$$g(K, p, w) = pF(K, L^*(K, p, w)) - wL^*(K, p, w) \tag{7}$$

giving maximal short-run profits as a function of the capital input, the output price, and the wage rate.

Before proceeding, we note some facts which will be used in what follows. The envelope theorem[2] implies that the derivative of $g(K, p, w)$ with respect to the output price is the short-run supply function, which we shall denote by $Q^*$,

$$\frac{\partial g}{\partial p}(K, p, w) = F(K, L^*(K, p, w)) = Q^*(K, p, w) \tag{8}$$

Similarly,

$$\frac{\partial g}{\partial K}(K, p, w) = pF_1(K, L^*(K, p, w)) \tag{9}$$

Moreover, differentiating (5) implicitly with respect to $K$ and $p$ gives

$$\frac{\partial L^*}{\partial K}(K, p, w) = -F_{12}/F_{22} \tag{10}$$

and

$$\frac{\partial L^*}{\partial p}(K, p, w) = -F_2/(pF_{22}) \tag{11}$$

If the firm hires its labor optimally in the short run, its long-run profits are

$$\pi(K, p, w, c) = pF(K, L^*) - wL^* - cK = g(K, p, w) - cK \tag{12}$$

At the time the capital input is chosen, $\pi$ is clearly a random function since $p$ is then a subjective random variable. In the following section it is assumed that the firm chooses its capital input to maximize the expected value of long-run profits. In Section IV we generalize by assuming the firm maximizes the expected utility of profits for some strictly concave von Neumann-Morgenstern utility function.

## II. The Risk-Neutral Case

In this section we assume the firm is risk neutral in the sense that it chooses the capital input to maximize the expected value of long-run profits,

$$Eg(K, p, w) - cK \tag{13}$$

An obvious implication of this assumption is that all uncertainty results we derive in this section arise from the non-linear nature of the technology rather than from risk aversion.

The first-order condition for the optimal capital input is

$$E\frac{\partial g}{\partial K}(K, p, w) = c \tag{14}$$

and the second-order condition is satisfied since $F$ is strictly concave. Indeed, using (9) and (10) we have

$$E\frac{\partial^2 g}{\partial K^2} = E\left[pF_{11} + pF_{12}\frac{\partial L^*}{\partial K}\right] = E[p(F_{11}F_{22} - F_{12}^2)/F_{22}] \tag{15}$$

Using (3), this clearly implies

[2] Paul Samuelson, pp. 34–35, gives a good discussion of the envelope theorem.

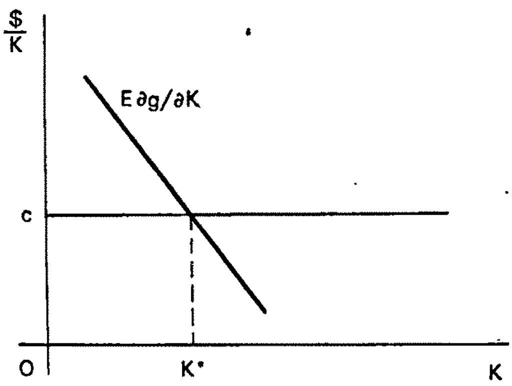



FIGURE 1

$$(16)\qquad E\frac{\partial^2 g}{\partial K^2}(K, p, w) < 0$$

The optimal capital input $K^*$ is shown graphically in Figure 1.

We now examine the effects of increased uncertainty on the demand for the capital input. The notion of increased uncertainty used in this section is the mean preserving spread recently studied by Michael Rothschild and Joseph Stiglitz. Operationally, a mean preserving spread is obtained by adding a random variable with conditional mean zero to the original random variable. Thus, we shall be examining the effects of replacing the random variable $p$ by $p+Z$ where $Z$ is a random variable with $E(Z|p)=0$. The notion of a mean preserving spread clearly includes both moving from one uncertain situation to another which is more uncertain as well as moving from a certain situation to an uncertain one. The result used is that the expected value of a convex (concave) function increases (decreases) as its argument undergoes a mean preserving spread. (See Rothschild and Stiglitz, p. 237.)

We determine the effects of increased output price uncertainty on the capital input by examining the behavior of $E\partial g/\partial K$ as $p$ undergoes a mean preserving spread. The Rothschild and Stiglitz results imply that if $\partial g/\partial K$ is convex (concave) in $p$, the mean preserving spread increases (decreases) $E\partial g/\partial K$ for any value of $K$. From Figure 1 it is clear that any increase (decrease) in $E\partial g/\partial K$ leads to an increase (decrease) in the optimal capital input.

To determine conditions under which $\partial g/\partial K$ is convex or concave in $p$ we differentiate twice. Using (9) and (11), we find,

$$(17)\qquad \frac{\partial^2 g}{\partial K \partial p} = F_1 + pF_{12}\partial L^*/\partial p = F_1 - F_2\frac{F_{12}}{F_{22}}$$

$$(18)\qquad \frac{\partial^3 g}{\partial K \partial p^2} = -F_2\frac{\partial(F_{12}/F_{22})}{\partial L}\frac{\partial L^*}{\partial p}$$

Since $\partial L^*/\partial p > 0$, it follows that

$$(19)\qquad \frac{\partial^3 g}{\partial K \partial p^2} \gtrless 0 \quad \text{as} \quad \frac{\partial(F_{12}/F_{22})}{\partial L} \lessgtr 0$$

An equivalent set of conditions can be derived in terms of the short-run supply curve, equation (8). Indeed, differentiation gives

$$(20)\qquad \frac{\partial^2 g}{\partial K \partial p} = \frac{\partial}{\partial K}\left(\frac{\partial g}{\partial p}\right) = \frac{\partial Q^*}{\partial K}(K, p, w)$$

$$(21)\qquad \frac{\partial^3 g}{\partial K \partial p^2} = \frac{\partial^2 Q^*}{\partial K \partial p}(K, p, w)$$

Thus

$$(22)\qquad \frac{\partial^3 g}{\partial K \partial p^2} \gtrless 0 \quad \text{as} \quad \frac{\partial^2 Q^*}{\partial K \partial p} \gtrless 0$$

Since $\partial^3 g/\partial K\partial p^2 > 0$ implies $\partial g/\partial K$ is convex in $p$, it is clear that increased output price uncertainty increases $E\partial g/\partial K$ and therefore increases the optimal capital input provided that $\partial(F_{12}/F_{22})/\partial L < 0$, or, equivalently, provided that $\partial^2 Q^*/\partial K\partial p > 0$. Similarly, increased output price uncertainty decreases $E\partial g/\partial K$ and therefore decreases the optimal capital input provided that $\partial(F_{12}/F_{22})/\partial L > 0$, or, equivalently, provided that $\partial^2 Q^*/\partial K\partial p < 0$. Since the labor input decision and therefore the output decision are made after the uncertainty is resolved, the increased uncertainty will affect these decisions only through its effect on $K$.

In view of (10), it is clear that for any given output price, a larger capital input will increase or decrease the demand for labor depending on the sign of $F_{12}$. There is nothing in our assumptions about the production

function, however, to give $F_{12}$ a determinate sign. Thus, for any $p$ realized *ex post*, greater uncertainty in the distribution of $p$ *ex ante* may increase or decrease the demand for labor depending on whether the increased uncertainty increases or decreases the capital input and also on the sign of $F_{12}$.

Recall that with expected profit maximization, the assumption that all inputs are chosen before the price is observed implies that uncertainty has no effect on the output or on the inputs. The fact that uncertainty plays a role in the model considered here is not surprising, though, since the firm is operating in a different environment. Indeed, if after the output price is observed, the capital input turns out to have been a poor choice, the firm is able to compensate partially by suitably choosing the labor input. Obviously, the firm could not do this if the labor input were also chosen before the price is observed. If the firm can compensate for what *ex post* turns out to have been a poor decision, it seems intuitively clear that uncertainty (which is related to the size of the possible *ex post* error in the capital input) should affect the capital input decision and that this effect should depend, possibly among other things, on the ease with which capital and labor are substitutable. The example in the next section shows that the mathematical results derived above are consistent with these intuitive conclusions.

## III. An Example

To get some additional feeling for the content of the results in the previous section, let us examine the strictly concave $CES$ production function,

$$Q = [aK^{-\beta} + bL^{-\beta}]^{-\mu/\beta} \tag{23}$$

where $a>0, b>0, \beta>-1, \beta\neq 0$, and $0<\mu<1$. The parameter $\mu$ is a measure of returns to scale. The elasticity of substitution $\sigma$ is given by

$$\sigma = 1/(1+\beta) \tag{24}$$

Rather than examine $\partial(F_{12}/F_{22})/\partial L$, it is more convenient to consider $\partial(F_{22}/F_{12})/\partial L$. Obviously, the sign of $\partial(F_{22}/F_{12})/\partial L$ is opposite that of $\partial(F_{12}/F_{22})/\partial L$. Some rather tedious but straightforward manipulations give

$$\frac{F_{22}}{F_{12}} = \frac{b}{a}\left(\frac{\mu-1}{\mu+\beta}\right)\left(\frac{L}{K}\right)^{-(\beta+1)} - \frac{(\beta+1)}{(\mu+\beta)}\frac{K}{L} \tag{25}$$

Differentiating with respect to $L$, we find

$$\frac{\partial(F_{22}/F_{12})}{\partial L} = \frac{\beta+1}{\mu+\beta}\left[(1-\mu)\frac{b}{a}\frac{L^{-(\beta+2)}}{K^{-(\beta+1)}} + \frac{K}{L^2}\right] \tag{26}$$

Since $\mu<1$, the term in square brackets is positive as is $\beta+1$. Therefore $\partial(F_{22}/F_{12})/\partial L$ has the same sign as $\mu+\beta$. Substituting $\beta=(1/\sigma)-1$ we find that the sign of $\partial(F_{22}/F_{12})/\partial L$ is the same as the sign of $(\sigma\mu+1-\sigma)$. It follows that the sign of $\partial(F_{12}/F_{22})/\partial L$ is the same as the sign of $(\sigma-1-\mu\sigma)$. Therefore,

$$\frac{\partial(F_{12}/F_{22})}{\partial L} \gtrless 0 \qquad \text{as } \sigma \gtrless (1-\mu)^{-1} \tag{27}$$

This is illustrated in Figure 2. In region $A$, $\partial(F_{12}/F_{22})/\partial L>0$ and increased uncertainty in the output price decreases the optimal capital input. In region $B$, on the other hand, increased uncertainty increases the optimal capital input. This suggests that with sufficiently easy substitutability between capital and labor, increased uncertainty decreases the optimal capital input since little is lost by waiting until the uncertainty is resolved and hiring labor.

As the production function approaches constant returns to scale (as $\mu\rightarrow 1$), increased uncertainty leads to a larger capital input. This can be explained in terms of a fact noted by Walter Oi; namely, since the profit function is convex in $p$, increased uncertainty will increase its expected value. With constant returns to scale, both $L^*(K, p, w)$ and $g(K, p, w)$ are homogeneous of degree one in $K$. In this case $\partial g/\partial K$ is independent of $K$ and is, in fact, a per unit of capital profit function. Under these conditions $\partial g/\partial K$ is convex in $p$, and increased uncertainty increases its expected value. We have excluded the case of constant returns to scale

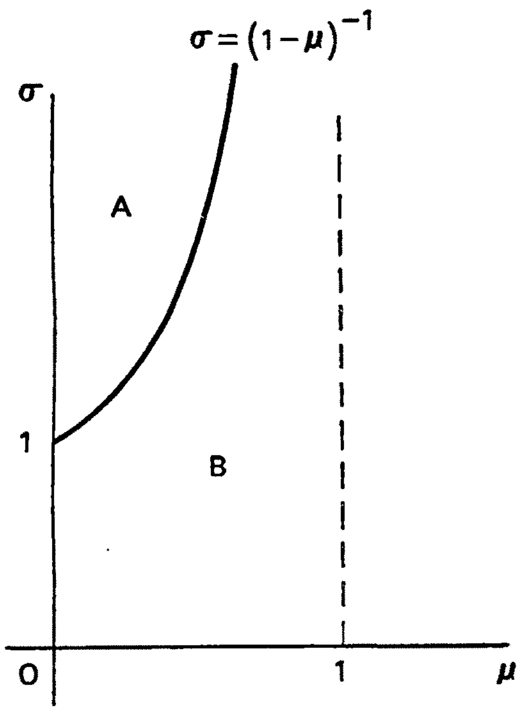



FIGURE 2

from consideration since the model considered here would not have determinate factor demand functions. However, when we are "close" to constant returns to scale we would expect increased uncertainty to have an effect on $E\partial g/\partial K$ similar to that with constant returns to scale.

As noted above, for any $p$ realized *ex post*, greater uncertainty in the distribution of $p$ *ex ante* may increase or decrease the demand for labor depending on the sign of $F_{12}$ and on whether the increased uncertainty increases or decreases the capital input. For the *CES* production function we are using as an example,

$$(28)\quad F_{12} = ab\mu(\mu+\beta)(aK^{-\beta}+bL^{-\beta})^{-(\mu/\beta+2)} \cdot (KL)^{-(\beta+1)}$$

$F_{12}$ clearly has the same sign as $(\mu+\beta)$, and this in turn has the same sign as $(\sigma\mu+1-\sigma)$. Thus,

$$(29)\quad F_{12} \gtrless 0 \quad \text{as} \quad \sigma \lessgtr (1-\mu)^{-1}$$

Figure 2 is again applicable. In region $A$, $\sigma > (1-\mu)^{-1}$ and $F_{12} < 0$. Now, in region $A$ increased uncertainty reduces the capital input, and since $F_{12} < 0$ the reduced capital input results in a larger labor input for any realized $p$. In region $B$, $\sigma < (1-\mu)^{-1}$ and $F_{12} > 0$. In this region the increased uncertainty increases the capital input, and since $F_{12} > 0$ the increased capital input results in a larger labor input for any realized $p$. For the *CES* example, increased output price uncertainty *ex ante* leads to a greater labor input *ex post* for any realized price except on the boundary between regions $A$ and $B$ where there is no change.

## IV. The Risk-Averse Case

In this section we assume the firm is risk averse. It maximizes the expected utility of profits,

$$(30)\quad EU(\pi) = EU[g(K, p, w) - cK]$$

where $U$ is a strictly concave von Neumann-Morgenstern utility function,

$$(31)\quad U'(\pi) > 0 \quad \text{and} \quad U''(\pi) < 0$$

Rather than examine the general case of a mean preserving spread, we simply compare the capital input in a certain situation with that in a situation involving a "small" amount of uncertainty.

Let $Ep = \bar{p}$, and let $\bar{K}$ be the optimal capital input if the price is known to be $\bar{p}$ with certainty. Thus $\bar{K}$ is the solution to

$$(32)\quad \frac{\partial g}{\partial K}(\bar{K}, \bar{p}, w) - c = 0$$

In the uncertain case, the firm maximizes (30). The first-order condition for this optimization is

$$(33)\quad EU'(\pi)\left[\frac{\partial g}{\partial K}(K, p, w) - c\right] = 0$$

The second-order condition is

$$(34)\quad EU''(\pi)(\partial g/\partial K - c)^2 + EU'(\pi)\partial^2 g/\partial K^2 < 0$$

and this is certainly satisfied since $U''(\pi) < 0$ and $\partial^2 g/\partial K^2 < 0$ by (16). The optimal capital

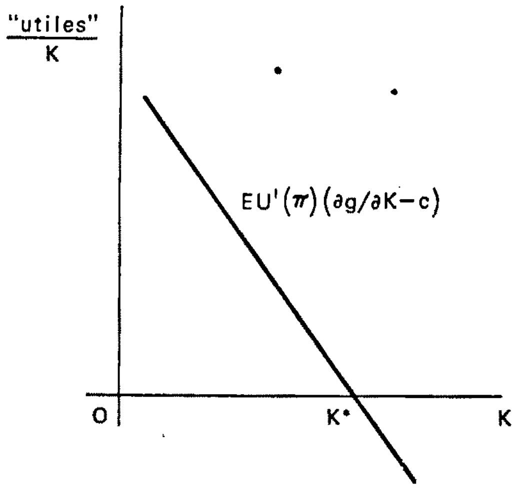



FIGURE 3

input with uncertainty, $K^*$, is shown graphically in Figure 3.

We now determine how $K^*$ and $\overline{K}$ differ. Suppose we insert the certainty capital input $\overline{K}$ in the left-hand side of (33). It is clear from Figure 3 that the certainty capital input is less than (greater than) the capital input with uncertainty if

$$(35)\quad E\left\{U'[g(\overline{K}, p, w) - \overline{K}c] \cdot \left[\frac{\partial g}{\partial K}(\overline{K}, p, w) - c\right]\right\}$$

is greater than (less than) zero. Our approach is to determine whether (35) is positive or negative.

Expanding the expression within braces in (35) around $p=\bar{p}$ and taking account of (32) gives

$$(36)\quad \left[U'(\pi)\frac{\partial^2 g}{\partial K \partial p}\right](p-\bar{p}) + \left[2U''(\pi)\frac{\partial g}{\partial p}\frac{\partial^2 g}{\partial K \partial p} + U'(\pi)\frac{\partial^3 g}{\partial K \partial p^2}\right]\frac{(p-\bar{p})^2}{2} + \dots$$

where all the terms in square brackets are evaluated at $(\overline{K}, \bar{p}, w)$. Taking the expectation of this expression, we find that (35) can be approximated by[3]

$$(37)\quad \left[2U''(\pi)\frac{\partial g}{\partial p}\frac{\partial^2 g}{\partial K \partial p} + U'(\pi)\frac{\partial^3 g}{\partial K \partial p^2}\right] \cdot \frac{Var\,(p)}{2}$$

where again the term in square brackets is evaluated at $(\overline{K}, \bar{p}, w)$ and where $Var\,(p)$ is the variance of the random variable $p$.

Now, by (8) $\partial g(\overline{K}, \bar{p}, w)/\partial p = Q^*(\overline{K}, \bar{p}, w)$ which is certainly positive. Differentiating again, $\partial^2 g/\partial K\partial p = \partial Q^*(\overline{K}, \bar{p}, w)/\partial K = (F_{22}F_1 - F_{12}F_2)/F_{22}$ by (9) and (11). None of our assumptions about the production function give this a determinate sign although one would generally expect it to be positive. This is certainly true for the *CES* production function considered above, as can be shown by some rather tedious calculations. Finally, $\partial^3 g/\partial K\partial p^2 = \partial^2 Q^*(\overline{K}, \bar{p}, w)/\partial K\partial p$ can be either positive or negative. In Section II we showed that it has the sign of $[-\partial(F_{12}/F_{22})/\partial L]$. (See equations (19) and (22).)

If $\partial Q^*/\partial K$ is positive and $\partial^2 Q^*/\partial K\partial p$ is negative then (37) is negative, and the presence of uncertainty reduces the optimal capital stock. In terms of the example involving the *CES* production function, this corresponds to being in region $A$ in Figure 2; and, as we showed in Section III, the smaller capital input due to uncertainty in this region leads to a larger labor input for any realized price. In the unlikely situation where $\partial Q^*/\partial K$ is negative and $\partial^2 Q^*/\partial K\partial p$ is positive, the uncertainty increases the optimal capital stock. In the rather likely situation where both $\partial Q^*/\partial K$ and $\partial^2 Q^*/\partial K\partial p$ are positive, the expression can be of either sign depending on the relative magnitudes of the terms. For the *CES* production function example, this corresponds to being in region $B$ in Figure 2. However, if $U''$ is sufficiently small in magnitude, the $\partial^2 Q^*/\partial K\partial p$ term will dominate and the capital input in the uncertain situation will be larger than in the

[3] Note that I have used the first-order condition under certainty, (32), to simplify the second-order term in (36). Thus, (37) is a valid approximation only for small amounts of uncertainty.

certain situation. If the capital input is larger in the uncertain case, we know from the analysis of Section III that for a *CES* production function this implies a larger labor input and a larger output for any realized price.

These results can be compared with those obtained when both inputs are chosen before the price is observed. In that case we have the Sandmo result, noted in the introduction, that the presence of uncertainty reduces output. Moreover, since the *CES* production function is homothetic, the reduction of output due to uncertainty will certainly reduce both of its factor demands.

## V. Concluding Comments

The results of Section II are particularly strong in that they include the effects of increasing uncertainty as well as the effects of moving from a certain to an uncertain situation. Since the firm is risk neutral in Section II, the uncertainty results derived there have nothing to do with risk aversion. They arise because the firm can partially compensate for what, *ex post*, appears to be a poor decision regarding the capital input by suitably choosing the labor input. I stress that the possibility of doing this is absent if all inputs must be chosen before the uncertainty is resolved.

The results of Section IV are much weaker in that we only compare a certain situation with one involving a small amount of uncertainty. The situation considered in Section IV is rather complicated since the uncertainty effects discussed in Section II frequently run counter to the uncertainty effects arising from risk aversion. However, the analysis does show that the effects discussed in Section II may well dominate the risk aversion effects in many cases, and thus lead to rather different results than those derived under the assumption that all inputs are chosen before the uncertainty is resolved.

## REFERENCES

**R. N. Batra and A. Ullah**, "Competitive Firm and the Theory of Input Demand Under Price Uncertainty," *J. Polit. Econ.*, May 1974, *82*, 537–48.

**R. Hartman**, "Competitive Firm and the Theory of Input Demand Under Price Uncertainty: Comment," *J. Polit. Econ.*, Dec. 1975, *83*, 1289–90.

**W. Oi**, "The Desirability of Price Instability Under Perfect Competition," *Econometrica*, Jan. 1961, *29*, 58–64.

**M. Rothschild and J. E. Stiglitz**, "Increasing Risk I: A Definition," *J. Econ. Theory*, Sept. 1970, *2*, 225–43.

**P. Samuelson**, *Foundations of Economic Analysis*, Cambridge 1947.

**A. Sandmo**, "Competitive Firm Under Price Uncertainty," *Amer. Econ. Rev.*, Mar. 1971, *61*, 65–73.

# On the Design of Managerial Incentive Structures in a Decentralized Planning Environment

*By* JOHN P. BONIN*

In a recent issue of this *Review*, Liang-Shing Fan presents a bonus scheme intended to induce a manager to report true performance capabilities to a planner. Such properties are extremely desirable for the decentralized planning procedures presently under discussion in several industrial socialist countries. Of current interest is a system called "planning from below," hereafter *PFB*, in which managers self-impose performance norms rather than receive targets from the planner (see Yevsey Liberman's discussion of the profitability norm). Incentives to induce a manager to report and strive for maximal performance are crucial in such a system. This paper considers informational and motivational properties of Fan's scheme and a similar piecewise linear bonus scheme in the decentralized *PFB* environment. For the general stochastic case, Fan's scheme is shown to yield somewhat peculiar norm-reporting results but to possess desirable motivational features.

In *PFB*, as in other decentralized planning environments, the planner interacts with the managers during the decision-making process. Information is initially dispersed so that detailed knowledge about any individual production technology is available only to the manager. The planner uses his knowledge of economic interdependencies and the overall goals of the system, along with any information he receives from the manager, to influence economic activity. The managerial incentive structures are designed in advance by the planner to coax the desired behavior. However, the distinguishing characteristic of *PFB* is that the manager imposes a performance norm upon himself which replaces planner-determined targets in evaluating and rewarding his actual performance. Consequently, the planner seeks to design incentive structures which motivate both norm prediction and performance representing the true capabilities of each manager.

In what follows, the scenario is a single planning period with intertemporal phenomena ignored. Input supply uncertainties make actual performance a random variable when the manager must determine his optimal norm. Although no input delivery plan is available prior to norm reporting, the planner may use the information from the norms to coordinate destinations and dates of delivery for intermediate commodities. Both managerial incentive schemes considered below are piecewise linear functions of actual performance with a kink discontinuity at the reported norm, i.e., the self-imposed target. The manager is assumed to be risk neutral so that he maximizes expected bonuses when calculating his optimal norm. The information and motivational results derived for both bonus schemes are not sensitive to any properties of the manager's probability distribution characterizing supply uncertainties.

In the next section, Fan's symmetric bonus scheme is shown to be preferable to other kink discontinuous ones which penalize deviation of performance from the target level (see Wayne Leeman). Section II considers a piecewise linear bonus scheme which allows nonsymmetric prediction error costs. This incentive scheme permits manipulation of relative prediction error costs so that a manager can be coaxed to take any amount of downside risk desired by the planner. Thus, the planner gains useful information about the expected fulfillment of norms and

* Assistant professor of economics, Wesleyan University. This research was undertaken while I was on sabbatical leave from Wesleyan University and visiting the University of California, San Diego. I am indebted to the former institution for its financial support and I wish to thank the latter for its hospitality and support. I have benefited from conversations with Paul Newbold and Clive Granger. I also wish to thank Alan Marcus for his helpful suggestions on an earlier draft.

can enforce some control on the *PFB* environment. The conclusion summarizes the major results and suggests a possible extension of this method to consider intertemporal phenomena in a different environment.

### I. Fan's Bonus Scheme

So that what follows may be interpreted in various institutional settings, the performance variable will be denoted $X$. In Fan's scheme $X$ would be profit, while in Leeman's it would be output. The manager's knowledge about technology and supply uncertainties is represented by a probability density function $f(X)$ defined for positive values of $X$, a random variable. Maximization of expected bonuses conditioned by $f(X)$ yields the optimal self-imposed target. In Fan's stochastic example, the norm reported is equal to the mean, median, and mode of a simple discrete probability distribution.[1] Hence, interpretation of normal capacity is simple and the incentive-unbiased result for norm reporting straightforward in both the deterministic and the stochastic case. However, when $f(X)$ is any arbitrary density function, we show the optimal norm under Fan's scheme to be the median of $f(X)$.

To derive Fan's optimal norm, rewrite his equation (2), p. 227, as

$$(1) \qquad B = \alpha(X - \epsilon|X - X_0|)$$

where $\alpha>0$, $1>\epsilon>0$, and $|\ |$ denotes absolute value.[2] With bonus written in this fashion, $\epsilon|X-X_0|$ can be interpreted as the nonnegative cost of a norm prediction error for any performance level $X$. It is clear why the manager will always report true capacity in the deterministic case first considered by Fan. Since the minimal value of $\epsilon|X-X_0|$ occurs when $X=X_0$, reporting and achieving true capacity $X'$ yields the maximum bonus $\alpha X'$. The same deterministic behavior is elicited from the bonus scheme suggested by Leeman, p. 436. Contrary to Leeman's assertion on p. 438, bonus schemes that reward overfulfillment of targets need not lead the manager to understate capacity. Indeed, Fan's scheme is such an example, at least in the deterministic case.

In most stochastic environments the definition of normal capacity is not obvious so that Leeman's statement is difficult to evaluate under uncertainty. Interestingly enough, Fan's scheme induces the manager to always report the median of $f(X)$ as the optimal norm. The median would surely be a possible candidate for normal capacity, therefore suggesting a stochastic counterexample to Leeman's assertion. Under uncertainty, expected bonus is written,

$$(2) \qquad E(B) = \alpha[E(X) - \epsilon E(|X - X_0|)]$$

The manager maximizes (2) with respect to $X_0$, given $f(X)$, to determine the optimal norm $X_0^*$. This is formally equivalent to minimizing, with respect to $X_0$,

$$(3) \qquad E(|X - X_0|) \equiv \int_{X_0}^{\infty} (X - X_0)f(X)dX - \int_0^{X_0} (X - X_0)f(X)dX$$

Differentiating (3) with respect to $X_0$ and solving the first-order condition for a minimum yields

$$(4) \qquad F(X_0^*) = \tfrac{1}{2}$$

where $F(X)$ is defined to be the cumulative distribution of $X$. The second derivative of (3) is $2f(X_0)$, which is always positive. Hence, $X_0^*$ determined by (4) minimizes (3). Now $X_0^*$ is simply the median of $f(X)$ by definition. The optimal norm for the Fan manager to report in any stochastic environment is the median of the probability distribution over performance values.[3]

[1] The distribution used by Fan is $f(X')=.5$, $f(X'+\delta)=.25$, and $f(X'-\delta)=.25$ where $X'$ is normal capacity.

[2] In Fan's equation (2), it is obvious that $\epsilon<1$ is envisioned. The rationale for such restriction surfaces when we examine the problem of performance concealment once a target is determined.

[3] The result in Fan's case, when the prediction error is a symmetric piecewise linear function with a single kink at the norm, is completely independent of characteristics of the probability distribution $f(X)$. Some care must be taken to conceptualize the median in discrete cases, but the median, suitably defined, is the optimum performance predictor for the entire class of probability distributions under Fan's scheme.

For purposes of comparison, it would be useful to know what norm the manager rewarded by Leeman's scheme would report in similar circumstances. Unfortunately, no general result can be shown. The various conditions on $f(X)$ under which a symmetric non-linear penalty function for prediction error yields simple norm predictors are given by Clive Granger, pp. 202–03. Although no obvious comparison can be made for norm reporting, Fan's scheme can be shown to be a considerable improvement over Leeman's in motivating performance.

An extremely desirable property of Fan's scheme is that for any *ex ante* prediction, bonuses are maximized at the *ex post* realized value of the performance variable, denoted $X'$.[4] Even if the planner cannot effectively monitor the performance variable, as would likely be the case with profits, the manager gains nothing by concealing performance capabilities. If the predicted norm is not achievable, i.e., $X' < X_0^*$, the maximum bonus is acquired by reporting $X'$ and receiving payment of $\alpha[X' - \epsilon(X_0^* - X')]$. Any other $X < X'$ would generate less bonus. If capacity is initially underestimated, i.e., $X' > X_0^*$, the maximum bonus in Fan's scheme is $\alpha[X' - \epsilon(X' - X_0^*)]$, as long as $\epsilon < 1$. Any other feasible performance $X < X'$ would yield a lower bonus, since the reduction in prediction error cost is less than proportionate to the reduction in $X$.

In contrast, for Leeman's scheme, any initial underestimation of capacity would induce the manager to conceal the higher realized performance, since bonus for performance at norm level exceeds bonus at above-norm levels. This property is most undesirable when the planner cannot easily monitor performance. Because Fan's scheme rewards for overfulfillment, no such motivational problems occur. Rather, the major role of the norm is to convey information to the planner about expected performance. Would the planner prefer to know some statistic other than the median for the purpose of coordinating intermediate transactions? If so, can a bonus structure be formulated which possesses the desirable motivational properties yet elicits different information? The next section addresses the issue of designing optimal bonus structures in *PFB*.

## II. The Planner's Strategy

In a different planning environment, Michael Keren has analyzed the interaction of planning agents when managerial bonuses depend on the fulfillment of a target set by the planner. Crucial to Keren's results is the characterization of optimal managerial performance as a simple function of the imposed output target. Bonus for overfulfillment would severely complicate his analysis. In *PFB*, the target is self-imposed and bonus for overfulfillment is crucial for stimulating performance. In what follows, we attempt to capture the interaction of agents in the *PFB* environment without sacrificing desirable motivational properties.

A piecewise linear bonus scheme with nonsymmetric prediction error cost is considered. By adjusting the weights applied to penalties for misestimating performance, the planner can coax the manager to report a safe or a strenuous norm. By differentiating managerial bonus structures across firms, the planner can impose any desired degree of risk on the overall fulfillment of reported targets. If output norms are reported, the resulting self-imposed production plan can be tuned to any level of tautness. The planner knows in advance the probability that each manager attaches to the attainment of his reported norm. Thus, the planner's risk preference is effectively transmitted to the manager through the bonus scheme with no consequent motivational problems in norm reporting or performance. The planner's strategy is more subtle in *PFB* situations, but the manager is nonetheless manipulatable by judicious design of the incentive structure.

If symmetric costs of prediction error are not imposed, the piecewise linear bonus structure is written

[4] Fan asserts that managers "will strive to attain their target profits set by the planning authority" (p. 226). What becomes clear is that the target in Fan's scheme has no influence on performance. For this reason, I have considered the problem in the *PFB* environment. Indeed, the planner seems to have no role in Fan's scheme. Interestingly, Fan's assertion is correct for Leeman's scheme!

$$(5) \quad B = \begin{cases} \alpha[X - a(X - X_0)] \text{ when } X \geqq X_0 \\ \alpha[X + b(X - X_0)] \text{ when } X \leqq X_0 \\ \text{where } \alpha > 0,\ 1 > a > 0 \\ \text{and} \qquad 1 > b > 0 \end{cases}$$

Then,

$$(6) \quad E(B) = \alpha E(X) - \alpha\left[a\int_{X_0}^{\infty}(X - X_0)f(X)dX - b\int_{0}^{X_0}(X - X_0)f(X)dX\right]$$

Maximizing (6) with respect to $X_0$ is equivalent to minimizing the term in square brackets. The optimal norm is determined as a function of $a$ and $b$. In particular, with $F(X)$ as previously defined,

$$(7) \qquad F(X_0^*) = \frac{a}{a+b}$$

Now, as $a \gtreqless b$, the optimal norm, $X_0^* \gtreqless X''$, the median of $f(X)$. The planner can influence the norm reported by weighting the prediction error costs differently.

Intuitively, when $a$ is large relative to $b$, more serious penalties in terms of foregone earnings are incurred for misestimating the norm on the low side. Consequently, the manager accepts more downside risk and reports a norm which is strenuous compared to that reported in the symmetric case. However, the restrictions on the coefficients indicate that bonus still increases with target overfulfillment. Consequently, given any self-imposed target, the manager will always report the highest possible level of performance, i.e., the realized value of $X$. Thus, this generalized kink discontinuous piecewise linear bonus scheme is exploitable by the planner in *PFB*. A plan resulting from self-imposed targets can reflect any risk of underfulfillment desired by the planner. However, bonus is positively related to performance so that no incentive to conceal above-norm performance exists.

The usefulness of adjusting the coefficients in (5) is exemplified by considering a class of density functions intuitively representative of *ex ante* supply uncertainties in the *PFB* economy. Let $f(X)$ be unimodal and negatively skewed so that $\overline{X}$ the mean $< X''$ the median $< X'''$ the mode.[5] Intuitively, $X'''$ is normal achievable capacity if expected input supplies are forthcoming. Based on previous experience, the manager calculates input estimates which have a high probability of being fulfilled. However, unexpected problems may develop which affect the quantity and timing of input supplies. Bottlenecks causing anything from slight inconvenience to disastrous consequences for the production capability of certain managers can occur with positive probability.

To adjust the riskiness of any plan for intermediate flows determined from reported norms, the planner may prefer some managers to self-impose targets different from the median. For density functions with the assumed skewness, Fan's manager understates normal (modal) capacity and overstates the expected mean value of performance, i.e., $\overline{X} < X_0^* < X'''$. For the nonsymmetric bonus structure in (5), the planner can coax a norm closer to the mode or the mean by judiciously adjusting the ratio $a/b \equiv r$. When $r > 1$, $X_0^* > X''$, and when $r < 1$, $X_0^* < X''$. As long as $r$ is sufficiently close to one, the optimal norm more closely approximates the mode in the first case and the mean in the second case.

An economic interpretation of the relative cost ratio $r$ is rather cumbersome. It is the ratio of the opportunity cost per unit of above-norm performance (in terms of bonus that would have been paid had performance been correctly estimated by the norm) to the penalty cost of each unit of below-norm performance. The planner by adjusting $r$ affects the norm reported. Consequently, the planner can induce the manager to accept any amount of downside risk by increasing the relative opportunity cost of above-norm performance.

The manipulatory property of the bonus scheme makes designing differentiated incentive structures an important planner's strategy in the decentralized *PFB* environment. Consider an example where $X$ represents

[5] Keren, p. 482, suggests that the same skewness is intuitively justifiable as a characterization of supply uncertainty in a centrally planned economy.

output level and the economic system consists of a group of interdependent firms. The planner attempts to formulate a consistent input delivery plan based on his knowledge of intermediate flows, perhaps estimated from previous input-output coefficients. At the same time, the planner has quantified the goals of the system as relative weights on final commodities. When input-output coefficients are known with certainty, Michael Manove describes an algorithm by which production units can be ranked according to the importance of their own product to aggregate final output. High priority firms often produce a commodity which is used extensively as an input throughout the system. Consequently, a shortage of this commodity would reverberate throughout the system. Firms producing final commodities are at the bottom of the ranking. One could conjecture that the fineness of the priority partition depends solely on the quality of the planner's information about interdependencies. Here I refer only to high and low priority firms to illustrate the way in which the planner can manipulate the self-imposed targets in *PFB*.

By setting $r$, the planner determines in advance the probability that the reported norm will be fulfilled or overfulfilled. By the definition of $r$ and (7), the optimal norm is determined from

$$F(X_0^*) = \frac{r}{r+1} \tag{8}$$

The probability that norm $X_0^*$ is at least fulfilled is then given by

$$1 - F(X_0^*) = 1 - \frac{r}{r+1} = \frac{1}{1+r} \tag{9}$$

By setting $r<1$ for high priority firms, the planner knows that the norm reported has a better than 50 percent chance of being met. Hence, the output reported would be a relatively safe estimate of the aggregate availability of that commodity for use as an input by other firms. On the other hand, $r>1$ induces the manager to take more downside risk. Hence, the norm reported will be underfulfilled more than half the time.

The planner may then use his knowledge of interdependencies to assign the differentiated $r$'s across priority classes. Managers of high priority firms could be induced to report relatively safe targets which in turn would help insure adequate input supplies. Low priority firm managers could be coaxed to impose relatively strenuous norms upon themselves. Moreover, the planner can impose his risk preference on the entire plan. Indeed, by increasing all $r$'s, the planner decreases the likelihood that each individual norm is fulfilled. Hence, such a self-imposed plan would be more risky. A comparison of the planner's strategies according to some average of all $r$'s weighted by priority class, call it $\bar{r}$, would index *PFB* "plans" according to tautness. Therefore, the optimal degree of plan tautness, common to the central planning literature, applies to decentralized *PFB* environments as well.

A complete analysis of plan tautness in *PFB* requires a theory explaining the manager's formation of his expectations about performance levels. Clearly, tauter plans affect the probability of obtaining inputs on time. Supply uncertainties were the primary basis for formulating a probability distribution about performance levels. Therefore, any information a manager might obtain about the expected degree of overall tautness $\bar{r}$ would be used to characterize $f(X)$. Given our assumptions about dispersion of information and the type of vertical communication between the manager and the planner, such information would be learned from past experience. Perhaps a sequential model with managerial subjective probabilities formed by adaptive expectations could adequately address this problem.

## III. Conclusion and Extension to Other Environments

In a decentralized informationally dispersed stochastic environment, the piecewise linear bonus scheme with a kink discontinuity at the target possesses desirable properties. Fan's scheme, a special case in which costs of prediction error are symmetric, yields maximum feasible performance for any target since it rewards overfulfillment. At the same time, the manager reports neither a strenuous nor a safe norm, the latter con-

jectured by Leeman to result when bonus for overfulfillment is paid. Instead, for any probability distribution over performance, the manager reports the median. In contrast, Leeman's bonus scheme requires more structure on the probability density function to yield simple norm-reporting results, but more importantly it induces concealment of any above-norm performance.

Interesting interactions between the planner and the manager arise when nonsymmetric prediction error costs are allowed. By adjusting the ratio of the opportunity cost for underestimating performance to the penalty cost for below-norm performance, denoted $r$, the planner can coax a safe or strenuous norm from any manager. In all cases, the incentive to maximize performance given any self-imposed target is retained. The nonsymmetric bonus scheme may be safely manipulated by the planner to impose his own risk preferences on an entire physical production plan determined by the managers themselves. Increases in all $r$'s, or perhaps an increase in an appropriately weighted index $\bar{r}$, indicate a tauter plan. In this case, the probability of managers achieving self-imposed output targets falls, with possible effects on the entire system's ability to balance intermediate transactions. Nonetheless, by judiciously designing the managerial incentive schemes according to his knowledge of economic interactions, the planner can coax managers to respond to the goals of the system in *PFB*.

Despite remarks to the contrary (see p. 227), Fan does not allow the infamous ratchet effect to influence norm reporting or actual performance. To introduce the ratchet effect requires explicit consideration of the interaction between the manager's performance and the planner's strategy for intertemporal target variation.[6] Both in *PFB* and in Fan's model, current norms and performance do not affect the probability of target fulfillment or overfulfillment in any future period. Such dependence led Keren to conjecture the importance of the ratchet effect when bonus is paid for overfulfillment in the jump discontinuous case (see Keren, p. 471, fn. 5). However, the structural similarity of the problem of norm reporting with point estimate prediction in stochastic decision making suggests a potentially fruitful line of attack on this more complex bonus structure in a different planning environment.

[6] Sam Gindin discusses the ratchet problem in a two-period deterministic situation for a special type of target variation. Intuitively, when the bonus structure is jump discontinuous at the target, the manager is more interested in safe targets.

## REFERENCES

**L.-S. Fan,** "On the Reward System," *Amer. Econ. Rev.*, Mar. 1975, *65*, 226–29.

**S. Gindin,** "A Model of the Soviet Firm," *Econ. Planning*, 1970, *10*, no. 3, 145–57.

**C. W. J. Granger,** "Prediction with a Generalized Cost of Error Function," *Oper. Res. Quart.*, 1969, *20*, no. 2, 199–207.

**M. Keren,** "On the Tautness of Plans," *Rev. Econ. Stud.*, Oct. 1972, *39*, 469–86.

**W. Leeman,** "Bonus Formulae and Soviet Managerial Performance," *Southern Econ. J.*, Apr. 1970, *36*, 434–45.

**Y. Liberman,** "The Plan, Profits and Bonuses," *Curr. Digest Soviet Press*, Oct. 1962, *14*, 13–15.

**M. Manove,** "Nonprice Rationing of Intermediate Goods in Centrally Planned Economies," *Econometrica*, Sept. 1973, *41*, 829–52.

# The Dynamics of Inflation in Latin America: Comment

By ROGER R. BETANCOURT*

In a recent article in this *Review*, a statistical analysis of aggregate data for sixteen Latin American countries leads Robert Vogel to the following conclusion: "The most important result of the present study for this controversy [monetarist-structuralist] is that a purely monetarist model, with no structuralist variables, reveals little heterogeneity among Latin American countries, in spite of their extreme diversity" (p. 113). One purpose of this paper is to show that this inference is inappropriate given the results presented in Vogel's paper. Moreover, the appropriate inference of heterogeneity among Latin American countries sheds no light on the monetarist-structuralist controversy. For no monetarist would argue that the same monetarist model must apply to every Latin American country. Instead of dwelling on the controversy, this note focuses on answering two of the questions which arise from the appropriate interpretation of Vogel's results: given the monetarist model used by Vogel, are there patterns in the differences of behavior among the Latin American countries? Since the answer to this question is yes, then: what do these patterns indicate about the nature of the inflationary process in Latin America?

## I

In order to facilitate the exposition a brief review of the underlying model, equation (6) in Vogel's paper, is necessary. This model relates the rate of inflation $P'_t$ to the rate of change in the money supply ($M'_t$), the rate of growth of income ($Y'_t$), and the lagged change in the rate of inflation ($P'_{t-1}-P'_{t-2}$) as follows:

$$(1) \quad P'_t = k + dM'_t + eM'_{t-1} - (1+a)Y'_t + b(P'_{t-1} - P'_{t-2})$$

where $k$ is expected to capture trends in velocity; $d$ and $e$ capture the current and lagged adjustment of prices to changes in the money supply, respectively; $a$ indicates the effect of real income on the ratio of real balances held to real income; and $b$ captures the effect of the expected costs of holding real balances.

Vogel arrives at the conclusion previously quoted by applying the above monetarist model to all Latin American countries. He estimates the model introducing country dummies for all the slope coefficients. His own $F$-test rejects the null hypothesis that the same model (same slope coefficients) applies to all Latin American countries at the 1 percent level. But the author proceeds to ignore this result because the test only "barely" rejects the null hypothesis! Since the probability of type I error (rejecting the null hypothesis when it is true) is only 1 percent whereas the probability of type II error (accepting the null hypothesis when it is false) is likely to be much larger, the dismissal of this result is totally unwarranted on statistical grounds. Consequently, no substantive conclusion can be said to follow from the dismissal of this result. The appropriate conclusion which follows from the test is that there is heterogeneity among Latin American countries in the coefficients of the variables in the model.

## II

This section presents the result of attempts to find patterns in these heterogeneous results. A look at Table 3 of Vogel's paper, p. 108, which provides the individual country regressions, yields little information on this issue because the variability in the results makes it difficult to detect any patterns. Notwithstanding, Vogel's discussion as well as the basic data presented in his Table 1, p. 103, seem to suggest that the countries fall into three clearly identifiable groups and

* Associate professor of economics, University of Maryland. I thank R. Vogel for a listing of his data and L. Radey for her research assistance. The computer Science Center at the University of Maryland provided support for the computations.

it may be possible to identify patterns by comparing the between group results. The three clearly identifiable groups are: high inflation (*H.I.*) countries (Argentina, Uruguay, Brazil, Chile, and Bolivia); moderate inflation (*M.I.*) countries (Paraguay, Colombia, Peru, and Mexico); and low inflation (*L.I.*) countries (Nicaragua, Ecuador, Honduras, Costa Rica, Guatemala, Venezuela, and El Salvador).

This grouping provided the point of departure in my search for the patterns in the coefficients that the data supported. The null hypothesis that the slope coefficients were the same across all countries was rejected at the 1 percent level when tested against the alternative that the slope coefficients were the same within a group and were different between groups ($F^{8}_{231}=6.2461$). However, the null hypothesis that the coefficients only differ between groups was rejected at the 1 percent level when tested against the alternative that the slope coefficients differ for all countries ($F^{52}_{179}=2.1527$).

Since the data rejected the restriction that the slope coefficients differ only between groups, Argentina was taken out of the high inflation groups and its slope coefficients were allowed to be different.[1] The null hypothesis that the slope coefficients differ only between groups was rejected at the 1 percent level when tested against the alternative that the slope coefficients differ between the (new) groups plus Argentina ($F^{4}_{227}=11.276$). Moreover, the null hypothesis that the slope coefficients differ between the (new) groups plus Argentina could not be rejected at either the 1 percent or the 5 percent level when tested against the alternative that the slope coefficients are different for all countries ($F^{48}_{179}=1.3274$). Thus, the grouping modified to allow the coefficients for Argentina to be different yields the pattern that the data support.

## III

The results of the pooled regressions allowing the coefficients to differ between groups are presented in Table 1 for both the grouping which includes Argentina among the high inflation countries (second block) and the grouping which treats Argentina separately (first block).

Note that the main features of the results are the same in both parts of the table. First, the impact of increases in the rate of growth of the money supply, either current or lagged one year, on the rate of inflation is substantially larger in both high and moderate inflation countries than in low inflation countries. (The null hypothesis of no difference in the impact is rejected using a two sided $t$-test at the 10 percent level in three of the four cases in part *A* and in two of the four cases in part *B*.) Secondly, the impact of the rate of growth of income on the rate of inflation differs among the three groups of countries. For example, in the high inflation countries this coefficient is small, and not significantly different from zero at the 10 percent level. On the other hand, in the moderate inflation countries this coefficient is significantly different from zero but not significantly different from (minus) unity using a two-sided $t$-test at the 10 percent level whereas the same type of test in the low inflation countries yields the result that this coefficient is significantly different from both zero and (minus) unity. Finally, the coefficient of the lagged changes in the rate of inflation tends to be small and statistically insignificant. The only exception is the high inflation countries in part *A* of the table where the estimated coefficient is significantly different from zero at the 10 percent level.

One of the most striking results in Table 1 is the magnitude of the estimated coefficient for Argentina of the rate of growth of income. This estimate is much larger than the estimates for any of the groups as well as significantly different from zero at the 1 percent level. Since the individual country regressions indicated that the same might be

[1] Initially Venezuela was taken out of the low inflation group and its slope coefficients were allowed to be different. However, the null hypothesis that the slope coefficients differ only between groups when tested against the alternative that the slope coefficients differ between the (new) groups plus Venezuela yielded a very low $F$ ratio, i.e., $F^{4}_{227}=.08$. Both Argentina and Venezuela were chosen because they seemed to differ the most from the rest of the group in their socioeconomic characteristics.

TABLE 1—POOLED REGRESSION RESULTS[a]

| | $M'_t$ | $M'_{t-1}$ | $Y'_t$ | $P'_{t-1}-P'_{t-2}$ |
|---|---|---|---|---|
| **Part A** | | | | |
| $H.I.$[b] | .580** | .380** | .032 | .077* |
| | (17.61) | (10.9) | (.3) | (1.8) |
| $M.I.$ | .385** | .438** | −.776** | −.040 |
| | (3.1) | (3.2) | (3.4) | (.4) |
| $L.I.$ | .103 | .180* | −.317* | .001 |
| | (1.0) | (1.9) | (1.7) | (.0) |
| Argentina | .896** | .354* | −2.308** | −.051 |
| | (5.1) | (2.1) | (6.8) | (.7) |
| **Part B** | | | | |
| $H.I.$[c] | .585** | .375** | −.140 | .052 |
| | (16.7) | (11.7) | (1.4) | (1.3) |
| $M.I.$ | .375** | .432** | −.784** | −.040 |
| | (2.8) | (3.0) | (3.2) | (.4) |
| $L.I.$ | .097 | .171 | −.310* | −.001 |
| | (.9) | (1.6) | (1.7) | (.0) |
| $ALL$[d] | .586** | .407** | −.298** | .014 |
| | (17.2) | (11.0) | (3.1) | (.4) |

[a] One or two asterisks indicates that the coefficient is significantly different from zero at the 10 percent or 1 percent level, respectively, using a two-tailed test.

The first block presents the results for the pooled regression with common intercept allowing the slope coefficients to differ between groups and for Argentina. The intercept was positive and insignificant at the 10 percent level ($R^2=.8740$; $\bar{R}^2=.8675$).

The second block presents the results for the pooled regression with common intercept allowing the slope coefficients to differ between groups. Again the intercept was positive and insignificant at the 10 percent level ($R^2=.8516$, $\bar{R}^2=.8438$).

[b] The three groups are defined as in the text except for the high inflation group ($H.I.$) which excludes Argentina.

[c] The three groups are defined in the text.

[d] These are the results for the pooled regression with common intercept and not allowing the slope coefficients to differ. It is the same as one of Vogel's regressions and the intercept is negative and significantly different from zero ($R^2=.8195$, $\bar{R}^2=.8165$).

true for Uruguay, the pooled regression was run treating both Argentina and Uruguay separately. The estimate of the coefficient of the rate of income growth for Uruguay was also much larger than for any of the groups and significantly different from zero at the 1 percent level. Interestingly enough, the results for the three groups and Argentina were quite similar to those reported in Table 1.

These results indicate how the inflationary process in Latin America differs between the three groups of countries, for example, in the impact of the rate of growth of the money supply and the rate of growth of income. They also indicate how the inflationary process differs even within the high inflation group of countries, for example, in the impact of the rate of growth of income for Argentina and Uruguay. Moreover, these differences are not very sensitive to minor changes in specification. For instance, suppressing the lagged change in the rate of inflation leads to a new set of estimates for the pooled regressions such that in the twenty-one possible comparisons (taking all elements of the table together) the largest difference between the old and the new results is .037.[2] While more elaborate or alternative theoretical and econometric specifications will affect the magnitude of the point estimates, they are not likely to obliterate the differences; on the contrary, they are likely to introduce new sources of differences. Instead of searching for these additional differences in the context of more elaborate models,[3] I conclude by raising the question: what explains the differences already observed in the inflationary process between these groups of Latin American countries?

Answering the above question requires, in my opinion, a detailed comparative analysis of the countries in the different groups over the sample period. In particular, such an analysis might be able to ascertain whether the low inflation countries have refrained from using monetary policy consistently to promote growth or resolve social conflicts and whether the opposite is true for moderate and high inflation countries. It would also be of interest to ascertain if this difference in policies, assuming that it exists of course, is due to favorable external circumstances, conscious design, or the nature of the social, political, and economic environment prevailing in these countries. Finally, a detailed comparative analysis may also help explain

[2] Similarly, correcting for heteroskedasticity hardly affects the patterns in the point estimates.

[3] It doesn't follow that we consider this area of further research useless or unimportant; but, given the data limitations for cross-country analysis, the main elaborations worth exploring can't be implemented except for some individual countries; for example, see A. C. Pastore.

the association between the high rates of growth over the period and inflation in the moderate inflation countries.

## REFERENCES

A. C. Pastore, "Notes on the Recent Monetary Policy in Brazil," paper presented at Nat. Bur. Econ. Res. conference on *The Uses of Econometric Models in Latin America*, Mexico City, Nov. 1974.

R. Vogel, "The Dynamics of Inflation in Latin America, 1950–1969," *Amer. Econ. Rev.*, Mar. 1974, *64*, 102–14.

# The Dynamics of Inflation in Latin America: Comment

By EDMUND J. SHEEHEY*

In a recent issue of this *Review*, Robert Vogel extended to sixteen Latin American countries a technique used by Arnold Harberger to examine the chief causes of Chilean inflation. On the basis of pooled regressions for the sixteen countries, he concluded that with little heterogeneity, the different rates of inflation cannot be put to structural differences but must be attributed to varying rates of expansion of the money supply. Only monetary variables are employed and the link between inflation and the rate of money supply growth rests essentially on $\bar{R}^2$ so high as to leave little of the variation in the rate of price increase for other variables to explain. This note contends that in coming to such broad conclusions on the basis of only monetary variables, Vogel has committed an error that the Harberger model was designed to avoid. Some regressions based on Argentine data are presented by way of illustration.

As a partial justification for using only monetary variables, Vogel states the Harberger model is essentially monetarist. In his original study, however, Harberger focused specifically on the problem of multicollinearity between changes in the money supply and the wage rate. He suspected that the long history of Chilean inflation consisted of instances when increases in the quantity of money had brought rising prices in the absence of wage increases, other instances when wages had pushed up prices without being "financed" by expansion of the money supply, and still others when inflation had been accompanied by wages and prices rising together. Regression analysis cannot sort out these instances, he asserted, " . . . but it can give some indications of their relative importance, and can surely distinguish between the extreme positions that deny the explanatory power of either wage changes or monetary changes, once the other is taken into account" (p. 228). Thus Harberger contended that if either monetary changes or wage changes are excluded, the regression analysis might not be able to distinguish between these extreme positions.

A formulation such as Vogel's, using money supply changes and lagged values of the rate of inflation, can capture the variation in prices resulting from: 1) current and past instances in which money supply changes alone have driven up prices; 2) current and past instances in which monetary and structural variables have risen in step; and 3) past instances in which structural variables alone have been at work. The only instances that it clearly will not capture are those in which current structural variables have operated without accompanying money supply changes.

Vogel's high $\bar{R}^2$ indicate that these instances have been few relative to all others. However, because Vogel's model will attribute to monetary factors the variation in prices due to both current structural causes which have been financed by increases in the money supply and past causes which have affected the lagged price variable, these $\bar{R}^2$ cannot say anything about the relative importance of structural and monetary factors. The Harberger model is designed to identify the relative inflationary impact of structural and monetary factors when each has operated alone. If it is true that in the majority of instances both have operated together and that in an important number of instances, each has forced up prices on its own, then either a totally monetary model, such as Vogel's, or a totally structural one will leave only a small portion of the variation in prices unexplained. Their relative importance as independent inflationary forces will only show up in an equation which uses both sets of factors.

* Economic affairs officer, General Agreement on Tariffs and Trade. I assume sole responsibility for the views expressed here.

TABLE 1—REGRESSIONS ON ARGENTINE INFLATION, 1950–71

| | Intercept | $M'_t$ | $M'_{t-1}$ | $M'_{t-2}$ | $Y'_t$ | $P'_{t-1}$ | $EXR'_t$ | $W'_t$ | $W'_{t-1}$ | $\bar{R}^2$ | $SE$ | $D.W.$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (1) | 4.497 | 1.248 | .232 | −1.01 | −.590 | .264 | | | | .57 | 9.0 | 1.90 |
| | | (5.1) | (0.8) | (4.2) | (1.7) | (1.8) | | | | | | |
| (2) | −3.849 | | | | | | .338 | .377 | .523 | .78 | 6.5 | 1.75 |
| | | | | | | | (6.5) | (3.3) | (5.1) | | | |
| (3) | − .627 | .425 | .192 | − .604 | −.347 | .036 | .250 | .365 | .380 | .84 | 5.6 | 2.03 |
| | | (2.1) | (1.0) | (3.7) | (1.6) | (0.3) | (4.9) | (3.0) | (3.1) | | | |

*Sources:* See the author. With the exception of real supply for the years 1950 to 1966 which is taken from Diaz-Alejandro, p. 534, all data are from official publications of the Argentine Republic.
*Note:* All data are percentage changes (denoted by primes) in semiannual averages. The dependent variable is the wholesale price index; $M$ is currency plus demand deposits; $Y$ is real supply, defined as gross domestic product plus imports; $P_{t-1}$ is lagged wholesale prices; $EXR$ is the exchange rate, defined as total peso trade divided by total dollar trade; and $W$ is the wage rate. The $t$-values are given in parentheses below each coefficient; $\bar{R}^2$ is the multiple correlation coefficient adjusted for degrees of freedom; $SE$ is the standard error of the estimate; and $D.W.$ is the Durbin-Watson statistic. Beginning with the second half of 1950, there are 43 observations for each regression.

Based on Argentine data for the period 1950 to 1971, the regressions in Table 1 alternatively present extreme monetarist and structuralist models and one that combines the two. The dependent variable is the wholesale price index and all data are semiannual averages. In the first equation, a formulation similar to Vogel's would seem to reinforce his monetary hypothesis. Current and lagged changes in the money supply, the current change in real supplies, and the rate of inflation from the previous period explain almost 60 percent of the variation in the rate of growth of prices. The combination of positive and negative money supply coefficients supports the hypothesis of Adolpho Diz that the response in Argentina to money supply increases has been an initial overshooting with a sharp deceleration a year or more later. The Durbin-Watson statistic gives no indication that significant variables have been omitted.

Equation (2), however, employs only structural variables and confirms precisely the opposite view. The hypothesis underlying this formulation is that wage and exchange rate increases are the active inflationary agents and the money supply is passively increased in order to avoid a serious shortage of liquidity. While devaluations have their main effect through an almost immediate rise in the price of tradeable goods, the impact of wage increases is spread out over the following year.[1] All three coefficients are significant at well above the 5 percent level and most of the inflationary impact of wages comes with a lag of six months. The $\bar{R}^2$ is far higher than that of the monetary model and the Durbin-Watson statistic is in the acceptable range.

Equation (3) employs both sets of variables as a test of the relative importance of Harberger's three sets of instances. The coefficient for $M_t$ is still significant at about the 5 percent level, but both it and that for lagged prices are greatly reduced in size and significance. The performance of the structural variables argues that the inflationary pressure they have exerted on their own has been far more important than that of the money supply.

These results serve to reinforce Vogel's own evidence of errors in the specification of his model. They show that high $\bar{R}^2$ for a purely monetary formulation are by no means inconsistent with even higher ones for a purely structural model. For Argentina the use of monetary and structural variables together shows the course of rising prices as alternatively propelled by demand and cost pressures with the latter apparently playing the greater role. If Vogel is correct in concluding that the inflationary process in Latin America is fairly homogeneous, then the results for Argentina may well give an accurate description of that process.

[1] For a structuralist model of inflation that would justify this formulation, see Ronald Findlay. See also Carlos Diaz-Alejandro.

## REFERENCES

C. F. Diaz-Alejandro, *Essays on the Economic History of the Argentine Republic*, New Haven 1970.

A. C. Diz, "Money and Prices in Argentina, 1935–1962," in D. Meiselman, ed., *Varieties of Monetary Experience*, Chicago 1970, 111–22.

R. Findlay, *International Trade and Development Theory*, New York 1973.

A. C. Harberger, "The Dynamics of Inflation in Chile," in C. Christ, ed., *Measurement in Economics: Studies in Mathematical Economics in Honor of Yehuda Grunfeld*, Stanford 1963, 219–50.

E. J. Sheehey, "Stabilization Policy and Protective Discrimination in Argentina, 1967–1971," unpublished doctoral dissertation, Michigan State Univ. 1974; Univ. Microfilms, Ann Arbor.

R. C. Vogel, "The Dynamics of Inflation in Latin America, 1950–1969," *Amer. Econ. Rev.*, Mar. 1974, *64*, 102–14.

# The Dynamics of Inflation in Latin America: Reply

*By* ROBERT C. VOGEL*

Roger Betancourt and Edmund Sheehey have offered interesting comments on my recent article on inflation in Latin America. Since each of these comments has a rather different focus, they are dealt with independently, first Sheehey's comment and then Betancourt's. In discussing Betancourt's comment, predictions of inflation based on coefficients estimated in my earlier article are compared with actual rates of inflation in Latin America in the 1970's.

The empirical work in my earlier article focuses on extending the Harberger model to sixteen Latin American countries by means of pooled time-series cross-section regressions, while Sheehey's comment concentrates on a single country by essentially replicating the earlier empirical work of Carlos Diaz-Alejandro and Adolfo Diz on Argentina. In attempting to build a case for his statistical analysis, Sheehey misinterprets my article at several key points. First, my conclusions with respect to the monetarist-structuralist controversy do not rest essentially on high $\bar{R}^2$s, as Sheehey asserts, but rather on limited heterogeneity. This is clearly stated in my article, p. 113, and quoted by Betancourt, but recognition of this point would make Sheehey's statistical analysis largely irrelevant, since his analysis is concerned only with Argentina.

Second, the essentially monetarist nature of the Harberger model is, as Sheehey states, a partial justification for my omission of structural variables. However, Sheehey neglects to mention what is made very clear in my article, p. 106: the primary reason for not including wages and exchange rates is that comparable data do not exist for a sufficient number of Latin American countries.[1] To assemble and analyze such data could provide a major contribution to the literature on the monetarist-structuralist controversy, but it would also require extraordinary effort.

Third, Sheehey asserts that my article reaches broad conclusions on the monetarist-structuralist controversy. On the contrary, my article, p. 113, emphasizes the difficulty of reaching such conclusions by pointing out that even with roughly similar findings, Harberger (1963), Diaz-Alejandro (1965), and Diz reach quite different conclusions on the relative importance of monetary and structural causes of inflation. As indicated in my article, p. 113, and discussed below, a principal issue is whether wage rates or the money supply is the autonomous variable, and my suggestions for further research emphasize the importance of investigating this issue, as well as the potential usefulness of adding structural variables to the model.

Although Sheehey claims to be following Harberger's approach and criticizes my article for failing to do so, he in fact fails to follow Harberger in two very important respects. First, Sheehey's lagged price variable is not the explanatory variable used by Harberger or in my article (see pp. 105–06). Contrary to Sheehey's interpretation, my extension of the Harberger model does not involve using the lagged rate of inflation to explain the current rate of inflation. Harberger (1963, pp. 225–27) is very clear that it is the change in the expected rate of inflation which changes the expected cost of holding money and hence affects the rate of inflation. Accordingly, Harberger and I use the lagged change in the rate of inflation as an explanatory variable, whereas Sheehey uses simply the lagged rate of inflation.[2] Thus the equation estimated by Sheehey contains a lagged dependent variable and may be viewed as a

* Associate professor of economics, Southern Illinois University at Carbondale.

[1] Another reason given in my article, but also overlooked by Sheehey, is the rather limited importance of structural variables in the findings of Arnold Harberger (1963), Diaz-Alejandro (1965), and Diz.

[2] Diz, pp. 114–15, constructs a series for exogenous changes in inflationary expectations, while Diaz-Alejandro (1965) omits all lagged variables from his reported regressions.

lagged adjustment model.[3] This view gives a rather different interpretation to the coefficients of Sheehey's explanatory variables and, in particular, to the patterns of response over time.

Second, Harberger (1963, pp. 227–29) emphasizes the use of an autonomous wage rate as an explanatory variable, whereas Sheehey apparently uses a realized, *ex post* wage rate.[4] The importance of using an autonomous wage rate should be clear: the change in a realized, *ex post* wage rate will be closely correlated with inflation unless inflation has a significant impact on real wages, a position which is not widely held.[5] Sheehey's equation also includes the lagged change in the wage rate which, it might be argued, could not be influenced by the current rate of inflation, but such an argument overlooks the likely correlation between leads and lags of the same variable. In particular, Diaz-Alejandro (1970, p. 376) finds a correlation of .44 between the current change in the wage rate and the change in the wage rate lagged two (six-month) periods. Sheehey suggests that multicollinearity between changes in the money supply and changes in the wage rate may present a problem in determining the cause of inflation, but a more basic problem is which is the autonomous variable. Both Harberger (1963, p. 245) and Diaz-Alejandro (1965, pp. 123–24) explore the correlation between leads and lags in the money supply and wage rates in their attempts to resolve this issue, but Sheehey does not deal with this issue.

Since Diaz-Alejandro and Diz had previously applied the Harberger model to Argentina, Sheehey's comment on my article may be viewed as primarily a comment on these earlier studies. Curiously, however, Sheehey does not point out that his findings are strikingly different from those of Diaz-Alejandro (1965) and Diz. In contrast to the predominance of structural variables in Sheehey's findings, monetary variables are at least as important as structural variables in Diaz-Alejandro's (1965) results, and Diz finds that monetary variables are clearly dominant. These differences may be due to differences in the definitions of certain variables or to differences in the time periods covered.[6] However, this is primarily an issue between Sheehey and Diaz-Alejandro and Diz, and does not seem to warrant further discussion here.

Betancourt's comment makes what could have been a significant contribution to understanding patterns of inflationary behavior in Latin America. However, the findings from his empirical work are marred by a serious statistical flaw which will be discussed below. In addition, Betancourt misinterprets the conclusion of my article, p. 113. He is incorrect in asserting that I ignore the existence of heterogeneity in slope coefficients among Latin American countries, and the quote which he selects from my article to illustrate his point even states that there is heterogeneity. The point which my article does make in stating clearly that there is significant heterogeneity at the 1 percent level is that "the $\overline{R}^2$ obtained by allowing all coefficients to differ among countries is only .03 higher than the $\overline{R}^2$s obtained in the pooled regressions in Table 2" (p. 109). This result, that allowing for heterogeneity among Latin American countries brings such a relatively small improvement in the explanatory power of the Harberger model, seemed to me quite surprising at the time the article was written. Moreover, it still seems to me quite surprising considering the widely differing inflationary experiences of Latin American countries and the importance frequently attributed to structural differences.

Contrary to Betancourt's argument, my article never implies that a "monetarist would argue that the same monetarist model must apply to every Latin American coun-

[3] As pointed out by Marc Nerlove and Kenneth Wallis, Durbin-Watson statistics are no longer appropriate when lagged dependent variables are introduced.

[4] Sheehey in fact nowhere defines his wage variable, but since he cites Diaz-Alejandro (1970) as the source for some of his data, it can probably be assumed that he uses Diaz-Alejandro's wage variable.

[5] See, for example, Harberger (1964, pp. 325–29).

[6] For unexplained reasons, Sheehey extends his study two years beyond mine, but he does not go back before 1950, despite the fact that both Diaz-Alejandro (1965) and Diz do.

try." It is indeed clear that monetarists would not argue that it is *necessary* for all Latin American countries to have identical coefficients; for example, monetarists have found income elasticities of the demand for money which differ widely among countries. However, the lack of heterogeneity in a monetarist model could be a *sufficient* reason for rejecting the structuralist position and accepting the monetarist view. In short, Betancourt has confused necessary and sufficient conditions.

Despite his attack on the empirical work in my article, Betancourt has failed to follow proper statistical procedures in reaching his own conclusions. It would have been an important contribution if Betancourt had found three distinct groups of Latin American countries with homogeneous coefficients within each group.[7] Unfortunately, Betancourt's own $F$-tests reject this outcome at the 1 percent level of significance. However, he then goes on to experiment with the data without acknowledging that under such conditions tests of significance no longer have the same meaning.[8] On the basis of undefined socioeconomic characteristics, Betancourt separates Argentina and Venezuela from their respective groups, and he then puts Venezuela back in its group to achieve the results which best suit his purposes. Subsequently, Betancourt also separates out Uruguay along with Argentina because of their large negative coefficients for income. However, for reasons which he does not mention, he fails to separate out Colombia which has a larger negative coefficient for income than Uruguay. Given this experimentation with the data, the standard statistical tests are no longer applicable, and it is not clear what can be concluded about similarities and differences in inflationary behavior. Almost inevitably some groupings of Latin American countries will provide so-called significant patterns, but this is not the appropriate procedure for hypothesis testing.

The data generated by Latin American countries since 1969 present an unusual opportunity to test the applicability of the Harberger model under greatly changed circumstances. The 1970's, particularly 1973 and 1974, have brought dramatic increases in the rate of inflation in most Latin American countries, and these changes have been accompanied by lengthy discussions of imported inflation and various structural problems. It is thus important to inquire to what extent the Harberger model with coefficients estimated on the basis of data for the period 1950–69 can explain Latin American inflation in the 1970's. In particular, the predictive power of the coefficients based on pooled regressions can be compared with the predictive power of the coefficients based on individual country regressions and on Betancourt's grouping of Latin American countries.

The May 1976 issue of *International Financial Statistics* provides data for 70 of the 80 possible observations (sixteen countries over five years) for comparing actual and predicted inflation from 1970 through 1974. When the predicted rate of inflation is calculated using the coefficients from the individual country regressions reported in Table 3 of my article, the actual rate of inflation is underpredicted by more than 8 percent on the average, while the coefficients from Betancourt's groupings yield an average underprediction of more than 4 percent. On the other hand, the coefficients from the pooled regression (with common intercept) in Table 2 of my article underpredict the rate of inflation by less than 2 percent on the average. In terms of mean square error, the coefficients from the pooled regression and from Betancourt's groupings yield almost identical results, while the mean square error for the coefficients from the individual country regressions is almost twice as great. Thus the coefficients from the pooled regres-

[7] As Betancourt indicates, his division of Latin American countries into three groups is based on my categories, p. 102. However, Betancourt fails to mention that the patterns which he discusses in the third section of his comment had already been noted in my article, p. 108.

[8] Michael Lovell (1974) examines other aspects of such experimentation which he terms "data-grubbing." Lovell has also pointed out to me that the coefficients for lagged rates of inflation can be interpreted in the context of second-order difference equations (see Lovell, 1975, pp. 410–14); and examination of the coefficients in Tables 2, 3, and 4 of my article reveals no examples of explosive behavior.

sion perform somewhat better than the alternatives in predicting Latin American inflation in the 1970's, and it is particularly interesting to note that only the predictions based on the pooled regression do not substantially underpredict the recent upsurge of inflation in Latin America.

## REFERENCES


R. R. Betancourt, "The Dynamics of Inflation in Latin America: Comment," *Amer. Econ. Rev.*, Sept. 1976, *66*, 688–91.

C. F. Diaz-Alejandro, *Exchange-Rate Devaluation in a Semi-Industrialized Country: The Experience of Argentina 1955–1961*, Cambridge 1965.

———, *Essays on the Economic History of the Argentine Republic*, New Haven 1970.

A. C. Diz, "Money and Prices in Argentina, 1935–1962," in D. Meiselman, ed., *Varieties of Monetary Experience*, Chicago 1970, 111–23.

A. C. Harberger, "The Dynamics of Inflation in Chile," in C. Christ, ed., *Measurement in Economics: Studies in Mathematical Economics and Econometrics in Memory of Yehuda Grunfeld*, Stanford 1963, 219–50.

———, "Some Notes on Inflation," in W. Baer and I. Kerstenetzky, eds., *Inflation and Growth in Latin America*, Homewood 1964, 319–35.

M. C. Lovell, "Data-Grubbing," paper presented at the meetings of the Econometric Society, December 1974.

———, *Macroeconomics: Measurement, Theory, and Policy*, New York 1975.

M. Nerlove and K. F. Wallis, "Use of the Durbin-Watson Statistic in Inappropriate Situations," *Econometrica*, Jan. 1966, *34*, 235–38.

E. J. Sheehey, "The Dynamics of Inflation in Latin America: Comment," *Amer. Econ. Rev.*, Sept. 1976, *66*, 692–94.

R. C. Vogel, "The Dynamics of Inflation in Latin America, 1950–1969," *Amer. Econ. Rev.*, Mar. 1974, *64*, 102–14.

International Monetary Fund, *International Financial Statistics*, Washington, May 1976.

# The Relationship between Relative Prices and the General Price Level

*By* Daniel R. Vining, Jr. and Thomas C. Elwertowski*

"If one great group of commodities varies pretty uniformly in one direction, and another in a different direction (or even in the same direction but in a markedly different degree), then the task of restoring the level of prices can no longer be regarded as a purely objective *quaesitum*, a currency problem."
*F. Y. Edgeworth*

It is very nearly a truism among neoclassical economists that the public is wont to confuse changes in the general price level with relative price changes. Correspondingly, one might speak of the independence between the two as one of the central postulates of the neoclassical tradition.

William Stanley Jevons was among the first to identify this basic dichotomy in economic life as well as to attempt to explain the ordinary citizen's seeming incapacity to understand it:

It is almost impossible, too, for any person to detect the effects of the change in the standard in his own personal affairs. Besides the interference of changes in prices and profits and activity of trade, which affect most persons, each individual has his own fluctuations of prosperity and adversity. If his income is not variable, at least his expenditure is more or less so. There are a hundred items of expenditure, some increased, some diminished, and among the variety of unmeasured circumstances, it is impossible for him to *feel* the very slow and measured change of 20 or 30 percent, spread over twenty or thirty years. If his receipts are variable and increasing, he enjoys the consciousness of prosperity, and probably attributes it complacently to his own abilities and deserts. If his income consists of fixed dividends or rents, he receives the same pieces of money as before, and has no thought that they are not what they were. *It is when he comes to pay his household bills that he can alone feel the difference. And then the difference seems to arise from the deficient harvest, from the growth of population, from the extortion of tradesmen, from anything rather than the change of a British sovereign fresh from the Mint. Value is the most invisible and impalpable of ghosts, and comes and goes unthought of while the visible and dense matter remains as it was.* [p. 75, last italics added]

Jevons' diagnosis of this misperception as due to the very gradual rise in the cost of living, however, has been badly undermined by recent events: the ordinary citizen persists, in the face of rates of decline in his dollar's purchasing power of as much as 12 percent annually, in attributing this decline to the abnormal behavior of *certain* prices. How little progress has been made since Jevons' day on the part of the public towards accepting a revision of this view of inflation is suggested by the preoccupation of contemporary economists, in their public addresses and popular writings, with the explication of this simple dichotomy between relative prices and the general price level:

*Milton Friedman:* . . . what of oil and food to which every government official has pointed? Are they not the immediate cause of the price explosion? Not at all. It is essential to distinguish changes in *relative* prices from changes in *absolute* prices. The special conditions that drove up the prices of oil and food required purchasers to spend more on them, leaving less to spend on other items. Did that not force other prices to go down or to rise less rapidly than otherwise? Why should the *average* level of all prices be affected significantly by changes in the prices of some things relative to others? [1974, p. 73]

*Alan Greenspan:* I think that there is a general fiscal-financial-monetary cause of inflation which essentially overrides all other

* Regional science department, University of Pennsylvania. Acknowledgements are due to the University of Pennsylvania and to Carnegie-Mellon University for fellowship assistance. Thanks are also due to Robert Barro, Lawrence Klein, Robert Lucas, and Allan Meltzer for valuable suggestions and criticisms.

considerations. Put another way: I do not believe that, in the longer run, the general price level is determined by the sum of its parts. . . . The relationship between the price of steel, for example, and the price of wheat is a relative concept—that is, it is determined by the relative supplies and demands, and the absolute price level of each is not determined by that relationship, but by the aggregate price level. Algebraically, we know of course that if we define all relative prices—that is, the relationship of one price to all others—and state the total price level, all individual prices are algebraically calculable. I submit that is a bit more than a statistician's calculation; it is something not terribly different from the process by which our price system is determined. [p. 2]

*Edgar Fiedler:* Despite the seriousness of the inflation problem, it is not well understood. I am particularly conscious, for example, of the widespread confusion between inflation, which reflects the general level of prices, and price changes for individual commodities. Most people will point to the price increase for fuel oil or rice or scrap steel, each of which has doubled in the past year, and call it inflation. But that is no more valid than pointing to soybeans or pocket calculators, all of which can be bought today for less than last year, and calling it deflation. Thus, it is important to remember that inflation is what happens to all prices, on the average. When we forget that point we forget the fundamental nature of inflation—and then we start to think about corrective policies that tend to treat the symptoms of the problem rather than its cause. [pp. 2–3]

It is the thesis of this paper that the popular folklore in fact enjoys ample justification in the data; it shall be shown that the behavior of the general price level is related in a statistically systematic manner to the behavior of individual prices relative to each other. To establish this thesis, of course, we must first make more precise what is meant by the contrary notion that the general price level is independent of relative price behavior.

A recent essay in this *Review* by Robert Lucas provides us with the necessary formalization of this notion of independence to test it on the incomparably rich body of data available on prices in this country. The article is set apart from previous attempts in the neoclassical school to represent mathematically this dichotomy in that it allows relative prices to vary. Previous work (see, for example, Don Patinkin) depicted relative prices as constant and, therefore, since they never shift, obviously and trivially independent of changes in the general price level. It does not require an elaborate analysis of price data to know that the hypothesis of relative-general price independence in this form must fail. Lucas' innovation is to represent this independence in the language of probability theory: the movement in the general price level over time, the mean of the distribution of a population of individual prices in the Lucas model is independent of (in the sense that it has no effect on) the remaining parameters and form itself of this distribution; and in the latter is captured our notion of "relative price behavior." Unlike the form the hypothesis takes under Patinkin's hands, there is no a priori reason to expect (except for the persistence of popular feeling to the contrary) Lucas' hypothesis not to fit the facts; that is, that the two forms of variation, one of the general price level over time and the other of individual prices relative to each other at a particular time specified probabilistically will have any relationship.

To summarize Lucas' model, let

$$P_{it} = P_t + z_t \tag{1}$$

where $P_{it}$ is the logarithm of the price of the $i$th good at time t, $P_t$ is the logarithm of the general price level at time t and a random variable, and $z_t$ is also a random variable, but independent of $P_t$. The difference $P_{it}-P_t$, then, is the logarithm of the relative price of good $i$, and in contrast to Patinkin's model is a random variable rather than a constant. In the Lucas model, this random variable is assumed to be normally distributed with mean 0 and variance $\tau^2$ and to be the same for all goods, $i=1, \ldots, k$; and $P_t$ is likewise assumed to be normally distributed, but with mean $\overline{P}_t$ and variance $\sigma^2$. Thus, the variance in individual prices (we shall henceforth mean by "prices" their logarithmic values) around their mean $P_t$ is a constant, $\tau^2$, and is therefore independent of the degree of variability $\sigma^2$ in the general price level $P_t$ around its trend, $\overline{P}_t$. The form of the dis-

tribution of individual prices is likewise constant and unrelated to the behavior of $P_t$. In short, the familiar constancy of relative prices in neoclassical economics is translated into a constancy in the mode of variation in these prices.

It turns out that there is strong statistical evidence that the two parameters $\tau^2$ and $\sigma^2$ move together, that is, as $\sigma^2$ increases (as the general price level becomes more unstable or less predictable relative to its trend value $\bar{P}_t$), so does $\tau^2$ (the dispersion in relative prices increases), and vice versa. There is somewhat weaker evidence that the form of the distribution of individual prices itself varies with $\sigma^2$. To demonstrate this dependence of $\tau^2$ and of the form of the distribution upon the magnitude of $\sigma^2$, it is necessary to translate Lucas' assumption, (1), into a statement about the changes in relative and general prices from one period to the next, since the more accessible data on the prices of goods are those giving those goods' price *changes* (i.e., indices relative to some base year) rather than their actual values in dollars and cents per unit of merchandise. To do so, we proceed with a simple transformation of equation (1):

$$(2)\quad P_{i,t+1} - P_{i,t} = P_{t+1} - P_t + z_{t+1} - z_t$$

$$(3)\quad [(P_{i,t+1} - P_{i,t}) - (P_{t+1} - P_t)]^2 = (z_{t+1} - z_t)^2$$

The expected value of the left-hand term of (3) is the expected variance of the one-period changes in $P_{it}$, $i=1, 2, \ldots, k$, around their mean change $(P_{t+1}-P_t)$. Thus, if $E(z_{t+1}z_t)=0$, as it is assumed to be in the Lucas model,

$$(4)\quad \mathrm{Var}\,(P_{i,t+1} - P_{i,t}) = E(z_{t+1} - z_t)^2 = 2\tau^2 = \gamma^2$$

That is, the variance in the one-period changes in $P_{it}$ around their mean $(P_{t+1}-P_t)$ should be roughly constant; and, since constant, unrelated to the degree of variability in the mean change. Since for large $\sigma^2$, this latter variability will be correspondingly large; (4) in effect says that $\gamma^2$ is invariant under shifts in the parameter $\sigma^2$.

An analysis of the behavior of both wholesale and retail prices over the period 1948–74,

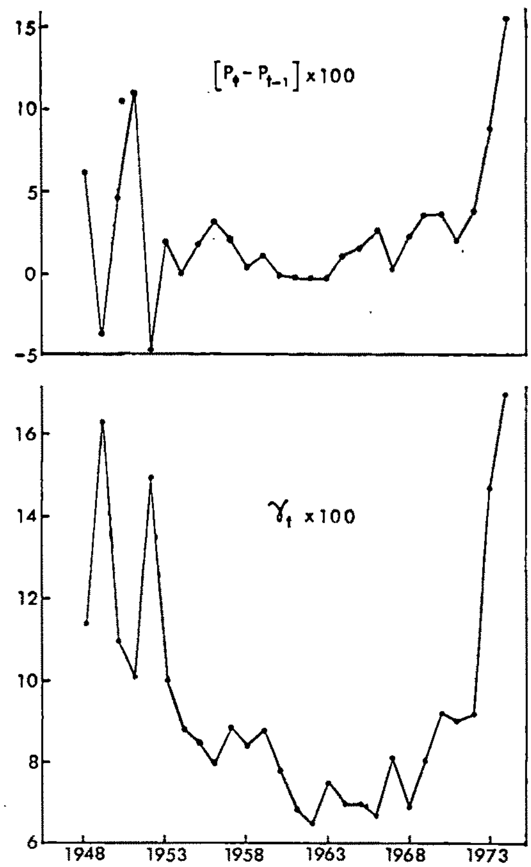



FIGURE 1. A comparison of $\gamma_t$ (the standard deviation in individual price changes at time t) and $(P_t-P_{t-1})$ (the average price change at time t) for the individual items of the Wholesale Price Index, 1948–74

however, reveals changes in $\gamma^2$ which are obviously in close association with the degree of general price change instability prevalent at time t, or $\sigma^2$. Consider first Figure 1, constructed from the components of the Wholesale Price Index (*WPI*). With the achievement of general price change stability in the late 1950's and early 1960's, the dispersion in individual price changes experiences a significant decline from its average level in the late 1940's and early 1950's, a period of great instability in the general rate of inflation. Subsequently, with what has come to be known as the Vietnam inflation, the standard deviation in these price changes rises to its original level—and beyond, with the energy and food scarcities of the early 1970's. Fur-

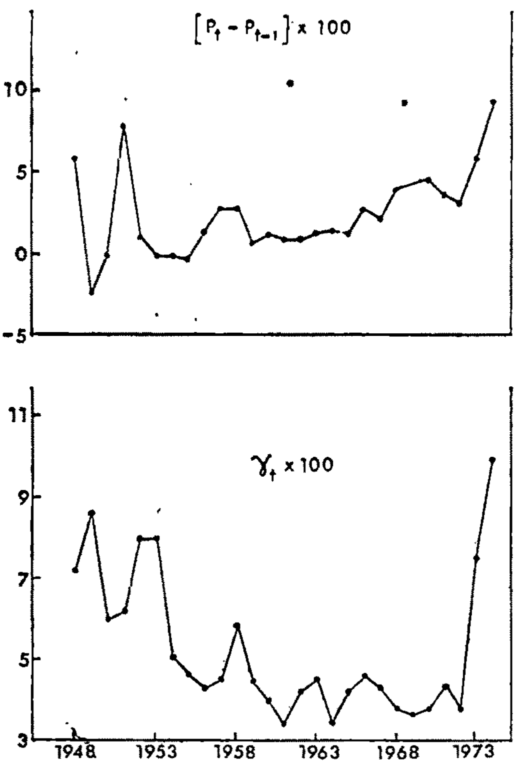



FIGURE 2. A comparison of $\gamma_t$ (the standard deviation in individual price changes at time t) and $(P_t - P_{t-1})$ (the average price change at time t) for the individual items of the Consumer Price Index, 1948–74

thermore, the transition from instability to stability in the inflation rate (i.e., from high to low $\sigma^2$), which is represented graphically as a gradual horizontal straightening in the line connecting the observations $P_{t+1}-P_t$, is characterized by a similarly gradual settling down of the dispersion in individual price changes $\gamma_t$.

The prices used to construct this figure are the price indices of the components of the *WPI* at the item (or 8 digit) level for each year between 1947 and 1974; the names and codes of these items are available from the authors. The basic data are available on tape (*WPI* Subfile 2002) from the Bureau of Labor Statistics (*BLS*). For any given year, the number of commodities for which a previous year's price index is available (and therefore for which $P_{i,t+1}-P_{i,t}$ may be calculated, as this difference is the logarithm of the ratio of the indices published by the *BLS*) varies between one and two thousand, with the smaller sample sizes at the beginning of the period of study and the larger at the end. All calculations have been performed on the unweighted first differences of the logarithms of the annual averages of these indices. Similar calculations were performed for a sample of commodities for which an unbroken series of price indices existed from 1947 to 1974 (795, in all); the resultant figure was found to be identical in form to that of Figure 1, although the individual values of $\gamma$ and $(P_{t+1}-P_t)$ for any given year tended to be somewhat different.

Figure 2 is identical to Figure 1, except that it has been constructed from the components of the Consumer Price Index (*CPI*) and thus represents retail rather than wholesale price behavior. Again, one notes high relative price change dispersion in the two episodes of general price change instability at the beginning and end of the period, 1948–74, and low relative price change dispersion in the one episode of general price change stability in the middle of this period. There does seem to be a delay relative to the behavior of wholesale prices in the rise of $\gamma_t$ with the onset of general price change instability in the late 1960's. The basic pattern, however, seems to be the same for both wholesale and retail prices.

The prices used to construct this figure are the price indices comprising the *CPI*, again at the individual item or 8 digit level and again available on tape from *BLS* (*CPI* Subfile 2001). The samples here are much smaller than those available for wholesale prices, varying between 110 in 1948 and 311 in 1973. As with our analysis of wholesale prices, calculations were performed on the unweighted first differences of the logarithms of the annual average price indices of those items for which two successive years' observations exist. We also performed these calculations on a sample of items for which an unbroken series of prices existed from 1947 to 1974 (there are 102 such items); the resultant figure is again identical in form to Figure 2.

Moreover, this relationship between general price change instability and relative price change dispersion seems to be a persistent feature of the American economy. As long ago as 1927, Frederick Mills, in his ex-

haustive study on wholesale price variation, observed that the dispersion in price changes in individual commodities increases with general or average price instability.[1] The reader is referred to page 284 of Mills where, after a long and fruitless attempt to relate this dispersion to a number of other aspects of general price behavior, such as its direction of change and even its absolute level, he uncovers a strong association between the absolute change in the general price level from one year to the next and the dispersion in relative price changes, or $|P_{t+1}-P_t|$ and $\gamma_{t+1}$. Since general price *change* instability (high $\sigma^2$) typically is closely associated with periods of general price instability (inflation or deflation, or high $|P_{t+1}-P_t|$), Mills' results and those represented graphically in Figures 1 and 2 are seen to be expressions of one and the same phenomenon.

In addition, the *shape* of the distribution of individual price changes, assumed in the Lucas article to be stable over time, is in fact highly variable and almost never of normal and infrequently of even symmetrical form. Rather, it is generally a highly skewed and asymmetrical distribution; and there are at least suggestions in the data that the direction of skew is the same as the direction of change in the rate of inflation (a high positive skew has been a particularly prominent feature of the current inflation). That is, if the general rate of inflation is climbing, then there tends to be a positive skew to the distribution, with most commodities' price changes below the mean (the rate of inflation) and a few above but generally at a great distance from it. On the other hand, if the general rate of inflation is falling, then the skew is typically a negative one with the majority of price changes greater than the mean and only a few below, but again generally at a great distance from this mean. The degree of skew is particularly pronounced during periods of rapid changes in the average rate of inflation. More stable periods, on the other hand, are characterized by distributions more nearly symmetrical. Their shapes, however, diverge from the normal in being much more peaked than would be expected if the samples were drawn from normal populations. (Tables 1 and 2 give the summary statistics, including those of kurtosis and skewness, for the retail and wholesale price change distributions of the years, 1948 through 1974.)

The latter feature, of course, is well known to students of speculative prices; indeed, that price changes, both speculative and otherwise, are distributed with significant amounts of kurtosis has been well known since the early teens of this century (see the papers by Benoit Mandelbrot and Wesley Mitchell). The presence of skewness in these distributions as well as their variability of form, and the suggestion of a systematic relationship between these two aspects of relative price variation and the mean of this distribution, i.e., the change in the general price level, has also been well documented, again by Mills.

The gradual settling down of the dispersion and form of the distribution of relative price changes with general price change stability (which corresponds here, and may always, to general price stability, i.e., constant $P_t$) may be what Friedman had tacitly in mind (although it seems to have been forgotten in the passage from his *Newsweek* column quoted above) when he recommended stable monetary growth on the grounds that it provides a stable background for the "other fluctuations": "It [the Federal Reserve] should limit itself to seeing to it that the stock of money grows at a steady rate month in and month out, year in and year out. This would provide a stable background

[1] Adopting Mills' approach and methodology, Frank D. Graham, in a book first published in 1930, uncovered a similar relationship between individual price change dispersion and general price instability during the German hyperinflation of the early 1920's (pp. 175–77). And an intercountry comparison of inflation rates and standard deviations in relative prices for 15 European countries by Herbert Glejser shows, again, an unmistakably strong relationship between the two. Glejser, however, seems to have employed *unlogged* price indices for different commodity groups (divided by the overall price index) as his basic observations. For lognormal variables, the standard deviation is known to have a positive relationship with the mean, though the variance in the logarithms of these variables remains stable. In this paper, we have shown the latter also to be positively related to the mean of the logarithms. We are indebted to Robert Barro and Lawrence Klein, respectively, for bringing our attention to these references.

TABLE 1—DESCRIPTIVE STATISTICS OF WHOLESALE PRICE CHANGE DISTRIBUTIONS, 1948–74

| Year | Mean[a] | Standard[b] Deviation | Skewness[c] | Kurtosis[d] | Sample Size |
|---|---|---|---|---|---|
| 1948 | .0625 | .1149 | −1.130 | 7.208 | 1159 |
| 1949 | −.0373 | .1636 | −2.331 | 8.457 | 1159 |
| 1950 | .0479 | .1090 | 2.342 | 16.727 | 1161 |
| 1951 | .1109 | .1013 | 2.010 | 10.663 | 1184 |
| 1952 | −.0483 | .1500 | −3.147 | 15.016 | 1183 |
| 1953 | .0020 | .1009 | −1.367 | 8.630 | 1182 |
| 1954 | −.0007 | .0883 | −0.410 | 8.751 | 1200 |
| 1955 | .0189 | .0852 | −0.427 | 9.622 | 1237 |
| 1956 | .0325 | .0798 | −0.273 | 10.103 | 1263 |
| 1957 | .0209 | .0890 | −1.880 | 26.161 | 1266 |
| 1958 | .0467 | .0843 | 0.605 | 29.350 | 1266 |
| 1959 | .0126 | .0902 | 0.122[e] | 11.761 | 1317 |
| 1960 | −.0014 | .0781 | −0.811 | 17.262 | 1359 |
| 1961 | −.0028 | .0683 | −0.432 | 6.933 | 1388 |
| 1962 | −.0039 | .0653 | −1.636 | 11.759 | 1522 |
| 1963 | −.0039 | .0751 | −0.300 | 30.929 | 1571 |
| 1964 | .0111 | .0697 | 1.259 | 14.538 | 1610 |
| 1965 | .0170 | .0699 | 0.580 | 12.266 | 1639 |
| 1966 | .0255 | .0668 | −0.247 | 11.126 | 1688 |
| 1967 | .0023 | .0817 | −2.284 | 13.153 | 1697 |
| 1968 | .0242 | .0685 | −0.158[e] | 9.776 | 1953 |
| 1969 | .0361 | .0806 | −0.162[e] | 25.862 | 1916 |
| 1970 | .0353 | .0925 | −0.311 | 15.303 | 1853 |
| 1971 | .0203 | .0905 | −1.295 | 9.898 | 1924 |
| 1972 | .0373 | .0915 | 2.602 | 22.040 | 1968 |
| 1973 | .0386 | .1471 | 2.541 | 7.107 | 2033 |
| 1974 | .1586 | .1707 | 1.198 | 5.680 | 1938 |

*Source: BLS, WPI Information Subfile 2002.*

[a] Let $P_{it}$ be the logarithm of the price index of commodity $i$ for year t, $\Delta P_{it}=P_{it}-P_{i,t-1}$, and $k$ the number of commodities. Then the mean price change $\Delta P_t$ is given by

$$\Delta P_t = \frac{1}{k}\sum_{i=1}^{k} \Delta P_{it}$$

[b] $$\gamma_t = \sqrt{\frac{1}{k}\sum_{i=1}^{k} (\Delta P_{it} - \Delta P_t)^2}$$

[c] $$\frac{\frac{1}{k}\sum_{i=1}^{k} (\Delta P_{it} - \Delta P_t)^3}{\gamma_t^3}$$

[d] $$\frac{\frac{1}{k}\sum_{i=1}^{k} (\Delta P_{it} - \Delta P_t)^4}{\gamma_t^4} - 3$$

[e] Indicates that a coefficient of skewness this large or larger could have been generated by a sample of size $k$ from a normal distribution with mean $\Delta P_t$ and standard deviation $\gamma_t$ with probability greater than .01. That is, only these observations are remotely consistent with the hypothesis of a normal distribution where the third moment is identically zero. None of the kurtosis statistics (again, their expected values are zero, if the underlying distribution being sampled from is normal) are consistent with a normal hypothesis. For details on normality tests using the third and fourth sample moments, see G. Snedecor and W. Cochran (1972).

for the other fluctuations" (Friedman, 1959, p. 623).[2] With stable growth in the money

[2] Raymond Bye interprets Mills' empirical discoveries explicitly in this fashion:

The proponents of plans for stabilizing the general price level are fully aware that different prices move up and down unevenly, in response to the myriad influences of demand and supply that affect particular commodities and they know better than to offer their proposals as a panacea which would correct all price disparities. This has been made perfectly clear in the writings of Irving Fisher and others. But the stabilizers rightly regard fluctuations in the level of prices arising from monetary changes as one serious cause of internal price disturbances. Mills' findings support rather than upset this view. [p. 47–48]

TABLE 2—DESCRIPTIVE STATISTICS OF RETAIL PRICE CHANGE DISTRIBUTIONS, 1948–74

| Year | Mean[a] | Standard[b] Deviation | Skewness[c] | Kurtosis[d] | Sample Size |
|---|---|---|---|---|---|
| 1948 | .0619 | .0729 | 0.741 | 2.496 | 110 |
| 1949 | −.0230 | .0863 | −1.095 | 2.307 | 110 |
| 1950 | .0000 | .0605 | 1.413 | 10.867 | 110 |
| 1951 | .0814 | .0623 | 0.995 | 3.629 | 111 |
| 1952 | .0122 | .0793 | 1.731 | 8.567 | 116 |
| 1953 | −.0042 | .0792 | −2.196 | 6.648 | 116 |
| 1954 | −.0013 | .0515 | −0.656 | 4.213 | 150 |
| 1955 | .0049 | .0469 | −0.413[e] | 5.741 | 150 |
| 1956 | .0142 | .0438 | −0.317[e] | 4.479 | 150 |
| 1957 | .0282 | .0455 | −0.728 | 9.278 | 151 |
| 1958 | .0298 | .0583 | 2.161 | 11.304 | 154 |
| 1959 | .0074 | .0440 | −1.396 | 4.983 | 154 |
| 1960 | .0123 | .0394 | −0.941 | 6.940 | 155 |
| 1961 | .0090 | .0348 | −0.168[e] | 3.199 | 159 |
| 1962 | .0099 | .0369 | 0.924 | 7.827 | 166 |
| 1963 | .0137 | .0459 | 4.098 | 26.843 | 167 |
| 1964 | .0143 | .0348 | 2.440 | 9.340 | 163 |
| 1965 | .0127 | .0424 | −0.341 | 9.835 | 286 |
| 1966 | .0290 | .0458 | −0.179[e] | 3.785 | 286 |
| 1967 | .0225 | .0432 | −0.970 | 5.805 | 303 |
| 1968 | .0408 | .0379 | 1.062 | 3.458 | 303 |
| 1969 | .0457 | .0363 | −0.473 | 2.794 | 303 |
| 1970 | .0469 | .0376 | 0.649 | 2.461 | 306 |
| 1971 | .0379 | .0429 | −0.945 | 3.923 | 309 |
| 1972 | .0310 | .0377 | 1.294 | 3.882 | 301 |
| 1973 | .0593 | .0750 | 2.234 | 5.195 | 311 |
| 1974 | .0958 | .0992 | 2.798 | 14.458 | 305 |

*Source: BLS, CPI Information Subfile 2001.*
a, b, c, d, e: See Table 1.

stock and, presumably as a consequence, a steady rate of inflation, the incessant fluctuations in relative prices normal to any developing economy with evolving tastes and technology are allowed to reach and maintain their "natural" level, i.e., they are no longer exaggerated by erratic movements in the monetary sector of the economy. It is interesting to note that what is perhaps the most appealing justification for a constant and unvarying expansion of the monetary base, namely, its elimination of artificially induced instability in the prices of products relative to each other, has yet to receive a formalization in economic theory: the link between the two kinds of instability, i.e., that of general and relative prices, remains a simple empirical regularity.

The increase in relative price change dispersion with general price change instability may also explain, in part, the hostility on the part of a large class of citizenry to price inflation, a resistance that some economists find irrational unless the inflation is accompanied by declines in real incomes, a comparatively rare event until recently. ("One man's price is another's income; when buyers pay more, sellers receive more. The inflation may proceed unevenly, so that some workers, consumers, and property owners lose while others gain; such relative distributional changes are always occurring, inflation or no inflation." James Tobin, p. 33.) If general price change instability is highly correlated with and accompanied by price inflation, and if dispersion in individual price changes widens in such periods, however, then a general atmosphere of higher risk and insecurity is the result, a cost of inflation in addition to that associated with the general depreciation of the currency over time which is the exclusive subject of Armen Alchian and

Reuben Kessel's study of its effects. That is, while the individual is more likely to experience a very large change in his price relative to that of all prices during periods of inflation, he is also more likely to experience a very small one relative to the average. The prevalence of positive skewness in such periods, whereby the majority of price changes are below the average price change (which may serve as a first approximation to changes in costs) would perhaps accentuate this preference for general price stability in a risk-averse population.[3] The widespread attraction of policy makers to price and wage controls may also be interpreted as a response, in some inchoate fashion, to the widening dispersion (and therefore riskiness) associated with inflationary periods. That is, wage and price controls may be a mechanism by which a normal variation in relative price changes is maintained *within* an environment of inflation and not a response to the general rate of inflation, as they are most often interpreted to be by economists (see, for example, James Barth and James Bennett, p. 397).

Indeed, one would hardly expect general price level instability to be the public policy problem that it manifestly is if it didn't have some systematic and predictable effect on the prices that are important to the individual, i.e., the prices of goods that he sells relative to those of the goods that he buys. A passage from J. M. Keynes (1932) makes this point in a particularly explicit way:

Money is only important for what it will procure. Thus, a change in the monetary unit, which is uniform in its operation and affects all transactions equally, has no consequence. If, by a change in the established standard of value, a man received and owned twice as much as he did before in payment for all rights and for all efforts, and if he also paid out twice as much money for all acquisitions and for all satisfactions, he would be wholly unaffected. It follows, therefore, that a change in the value of money, that is to say in the level of prices, is important only insofar as its incidence is unequal. Such changes have produced in the past, and are producing now, the vastest social consequences, because, as we all know, when the value of money changes, it does not change equally for all persons for all purposes. A man's receipts and his outgoings are not modified in one uniform proportion. Thus, a change in prices and rewards, as measured in money, generally affects different classes unequally, transfers wealth from one to another, bestows affluence here and embarassment there and redistributes Fortune's favors so as to frustrate design and disappoint expectation.[4] [pp. 80–81]

However, Keynes' insight seems generally to have been forgotten by the current generation of economists; thus, in contrast to many of his other conclusions, Lucas' remark in an otherwise extremely controversial paper, namely "that there is no reason to expect $\tau$ to vary systematically with demand policies" (1976, p. 39), has gone utterly uncontested.

It is perhaps at least worth remarking that this lack of awareness of the relative-general price relationship is even a contemporary phenomenon among neoclassical economists. As Lucas finds the *lack* of such a relationship entirely to be expected, his earlier counterparts in the neoclassical school declared its *existence* (specifically in their assessment of Mills' contribution) likewise unremarkable and quite consonant with the principles of

[3] A possible implication of this line of thought is that a falling rate of inflation would be more tolerable since the majority of price changes then would lie above the average price change; that is, there would be a greater than 50 percent chance that the rate of increase in the price of your good will exceed the average rate of increase for all goods. Richard Day has used this fact about negatively skewed distributions to explain the observed preference among Mississippi Delta farmers for oats over corn, though the average profitability of the latter exceeds the former. On the other hand, a pronounced skewness in either direction suggests an asymmetry in the distribution of the overall price change across its members that a society probably cannot long tolerate.

[4] Of course, as Tobin points out, "relative distributional changes are always occurring, inflation or no inflation" (p. 33). In not introducing the language of probability, which would have probably been inappropriate in this essay, Keynes constrains himself to write as if relative price change variation is zero in periods of general price stability. He would have been aware of the absurdity of this proposition, and perhaps intended it only as shorthand for the situation outlined here—namely that the variation in relative price changes is more violent in times of inflation.

that school:

> A significant correlation (+.614 for 35 pairs of observations) is found between degree of change—regardless of direction—in the price level and degree of dispersion of prices [page 284]. All that is necessary to adduce in explanation of this correlation is that, in addition to the many influences affecting particular commodities which operate to bring about dispersion of prices even when the price level is constant, active pervasive factors tending to affect all prices will further increase dispersion if different commodities have varying degrees of resistance to their influence. Jacob Viner [pp. 345–46]

Though there is something tautological about Viner's explanation that belies his nonchalance about the phenomenon Mills discovered, this was about as far as such explanations went among those aware of Mills' work. Raymond Bye, for example, ten years after Viner's review article, writes:

> Mills finds that it is the *amount* and not the direction of changes in the price level which affect the degree of internal stability in the price system. When the average movement of prices is violent in either an upward or downward direction, dispersion . . . become[s] greater. Apparently, marked changes in the price level, caused perhaps by monetary factors, seriously disturb the ordinary price relationships between different commodities. This could be inferred from general reasoning, because the incidence of monetary changes would hardly be instantly and evenly diffused through the system of prices. [pp. 37–38]

In fact, the first attempt to formalize, within the neoclassical tradition, an explanation of the phenomenon is in Robert Barro's extension of Lucas' 1973 article. Greatly simplified, Barro's explanation is that for a given random variation in the demands for the various goods making up the economy, a rotation in the supply curves for these goods towards greater inelasticity (brought about by the increased general or average price change instability, as in the Lucas model, where the individual suppliers become less price responsive as prices become more dominated by movements in the general price level) will generate a higher variability in the observed relative price changes. This may be grasped quite readily via the conventional supply-demand curve diagram. If the demand curve varies around some average curve, the steeper the supply curve the greater the variability in the equilibrium price for any given level of demand curve variation.

Barro's is a highly sophisticated effort to account for the observed dependence of heightened relative price change dispersion on general price change instability, relying upon a chain of causality running from general price level change instability to relative price change instability. It purposely ignores the perhaps more acceptable view, at least to the ordinary citizen, that an autonomous increase in relative price change dispersion somehow induces instability in the changes in the general price level. By this view, the latter is simply a symptom of the former, the former being the only item of interest to the citizen. The mechanism of causality might well be that of the neoclassicists, i.e., an increased instability in the money supply process. But the latter would have much less autonomy than is attributed to it by monetarists. This in fact is the tentative explanation put forward by Mills himself in his rejoinder as an alternative to the view of Bye:

> Bye's third principle ($MV+M'V'=PT$) is equally capable of a two-way cause and effect interpretation. Not necessarily is the monetary circulation antecedent, with price changes consequent. Under contemporary conditions, a change in the volume of credit outstanding may result from changes in the prices of individual commodities, changes that may in turn be due to any of an infinity of forces affecting demand or supply conditions. [See Bye, p. 139]

Mills' qualification here, "under contemporary conditions," is of interest, since we have been unable to find any relationship between relative price change dispersion and average price change instability in the data of Jevons (pp. 136–41) on commodity prices in England during the middle 1800's. Variations in the money supply during his day may very well have been autonomous, due as they were to the discovery of gold in remote parts of the world. Crises in the relative price

relationships could not therefore be "papered over" by increased issues of paper currency, leading in turn to a general rise in all prices. And the latter could not therefore in turn be interpreted, except under the contemporary conditions of a credit system under the control of the government, as simply the symptom of abnormal relative price disturbances.

A major challenge to empirical economists would seem to be the discrimination between these two hypotheses, i.e., to determine the direction of causality between individual price change dispersion and general price change instability. Both theories must have testable implications for the data that are available. The theory advanced by Barro may well have; it is at least in the form that one could answer one way or the other. The counter hypothesis to that of Barro has not acquired any more particulars since the original intuition of Mills, quoted above. This would seem to be the first order of business for economists interested in such matters.

## REFERENCES


**A. Alchian and R. Kessel,** "Effects of Inflation," *J. Polit. Econ.*, Dec. 1962, *70*, 521–37.

**R. Barro,** "Rational Expectations and the Role of Monetary Policy," *J. Monetary Econ.*, Jan. 1976, *2*, 1–32.

**J. Barth and J. Bennett,** "Cost-Push versus Demand-Pull Inflation," *J. Money, Credit, Banking*, Aug. 1975, *7*, 391–97.

**R. Bye,** *An Appraisal of Frederick C. Mills' The Behaviour of Prices* (with a rejoinder by F. C. Mills, a panel discussion, and a commentary by R. Bain), New York 1939.

**R. Day,** "Probability Distributions of Field Crop Yields," *J. Farm Econ.*, Aug. 1965, *47*, 713–41.

**F. Edgeworth,** "Measurement of Change in the Value of Money," in his *Papers Relating to Political Economy*, Vol. I, London 1925.

**E. Fiedler,** "Price Changes, Inflation, and Controls," presented to the Graduate School of Industrial Administration, Carnegie-Mellon University, Mar. 27, 1974; reprinted in *Department of Treasury News*, S-386.

**M. Friedman,** "Perspectives on Inflation," *Newsweek*, June 24, 1975, p. 73.

———, "Statement and Testimony," in U.S. Congress, Joint Economic Committee, *Employment, Growth, and Price Levels, Part 4*, "The Influence on Prices of Changes in the Supply of Money," Hearings, 86th Cong., 1st sess., Washington 1959, pp. 605–69.

**H. Glejser,** "Inflation, Productivity, and Relative Prices—A Statistical Study," *Rev. Econ. Statist.*, Feb. 1965, *47*, 76–80.

**F. D. Graham,** *Exchange, Prices, and Production in Hyper-Inflation: Germany, 1920–1923*, New York 1967.

**A. Greenspan,** "A General View of Inflation in the United States," in *Inflation in the United States: Causes and Consequences*, Proc. Conference Board Economic Forum, May 21, 1974, New York.

**W. S. Jevons,** *Investigations in Currency and Finance*, London 1909.

**J. M. Keynes,** "Social Consequences of Changes in the Value of Money," in his *Essays in Persuasion*, New York 1932.

**R. Lucas,** "Some International Evidence on Output-Inflation Tradeoffs," *Amer. Econ. Rev.*, June 1973, *63*, 326–34.

———, "Econometric Policy Evaluation: A Critique," *J. Monet. Econ.*, 1976, suppl., *2*, 19–46.

**B. Mandelbrot,** "The Variation of Certain Speculative Prices," *J. Bus., Univ. Chicago*, Oct. 1963, *36*, 394–419.

**F. C. Mills,** *The Behavior of Prices*, New York 1927.

**W. Mitchell,** "The Making and Using of Index Numbers," in *Introduction to Index Numbers and Wholesale Prices in the United States and Foreign Countries*, U.S. Bureau of Labor Statistics, Bull. 173, Washington 1915; reprinted as Bull. 284, 1921.

**D. Patinkin,** *Money, Interest, and Prices*, 2d ed., New York 1965.

**G. Snedecor and W. Cochran,** *Statistical Methods*, Ames 1972.

**J. Tobin,** "There are Three Types of Inflation: We Have Two," *New York Times*, Sept. 6, 1974, p. 33.

**J. Viner,** "Review of *The Behaviour of Prices*," *Quart. J. Econ.*, Feb. 1929, *43*, 337–52.

**U.S. Bureau of Labor Statistics,** CPI Information Subfile 2001, Washington 1975.

———, WPI Information Subfile 2002, Washington 1975.

# On the Shape of the Trade Indifference Curve: Rejoinder to Batra

*By* MURRAY C. KEMP AND EDWARD TOWER*

Raveendra Batra argues in his reply that our comment on his article is substantially incorrect because "the curvature of the trade indifference curve in the presence of the nontraded good is indeterminate" (p. 252). In this note we demonstrate the correctness of our earlier assertion that, even when there are nontraded goods, the trade indifference curve has the usual curvature.

Let $C=\{c=(c_1, c_2, \ldots, c_n)\in R^n_+ \mid U(c)\geq U_0\}$ be the production possibility set, where $T$ denotes the set of input-output combinations let $Y=\{y=(y_1, y_2, \ldots, y_n)\in R^n_+ \mid (y, \bar{x})\in T\}$ be the production possibility set, where $T$ denotes the set of input-output combinations that are feasible and $\bar{x}$ denotes the given factor endowments. Both $Y$ and $C$ are assumed to be convex. The set of trades which generate levels of utility not less than $U_0$ is simply $E=Y-C$, where $E=\{e=(e_1, e_2, \ldots, e_n)\in R^n\}$ and the $i$th good is exported or imported according as $e_i$ is greater than or less than zero. It is well known that any linear combination of two convex sets must also be convex.[1] Thus $E$ also is convex, and the surface which bounds it, a trade indifference surface when all goods are traded, must have the normal curvature.

Now in restricting our attention to the case in which $m$ goods are not traded, so that $c_i=y_i$, $i=n-m+1, \ldots, n$, we are considering that linear cross section of $E$ determined by $E\cap\Gamma$, where $\Gamma=\{e=(e_1, e_2, \ldots, e_n)\in R^n \mid e_i=0, i=n-m+1, \ldots, n\}$. Since $\Gamma$ is convex ($\Gamma$ is a linear subspace), $E\cap\Gamma$ must also be convex. Since this is simply the set circumscribed by the trade indifference surface when the last $m$ goods are nontradeables, we have proved that the trade indifference surface for an economy which consumes $n$ goods, $m$ of which are nontradeables $(m<n)$, has the normal curvature.

* Research professor of economics, University of New South Wales, and associate professor of economics, Duke University, respectively. We are grateful to Daniel Graham for helpful suggestions.

[1] See, for example, theorem 2.10 of H. Nikaido.

## REFERENCES


R. N. Batra, "Nontraded Goods, Factor Market Distortions, and the Gains from Trade: Reply," *Amer. Econ. Rev.*, Mar. 1975, *65*, 251–52.

M. C. Kemp and E. Tower, "Nontraded Goods, Factor Market Distortions, and the Gains from Trade: Comment," *Amer. Econ. Rev.*, Mar. 1975, *65*, 249–50.

H. Nikaido, *Convex Structures and Economic Theory*, New York and London 1968.

# The Terms of Trade and the Balance of Payments in the Short Run

By Carlos Alfredo Rodriguez*

Several recent models associated with the "monetary approach to the balance of payments" have focused on the role of adjustment to monetary disequilibrium as the crucial factor in explaining the behavior of the overall balance of payments.[1] It is the purpose of this paper to point out a peculiar characteristic of those models regarding the effects of an exogenous change in the price of a tradable good on the balance of payments. It will be shown that under the most likely assumptions, the simple versions of these models predict a short-run improvement in the balance of payments as a consequence of an exogenous increase in the price of any tradable good, whether exportable or importable. I also present an alternative formulation, still within the spirit of the monetary approach, which allows for different short-run effects of a price change on the balance of payments depending crucially on whether the commodity increasing in price is an exportable or an importable good. This alternative formulation, based on the "precautionary" demand for money, allows real cash balances to play the role of a short-run "shock absorber" in order to even out over time the required changes in the level of real expenditures.

## I

Modern work on the monetary approach to the balance of payments originated with Mundell and Johnson.[2] This approach correctly recognizes that in an open economy with fixed exchange rates, changes in the domestic supply of money can be brought about via domestic credit or foreign exchange operations by the monetary authority. This fact plus the assumption that the existing stock of money must be willingly held by individuals implies that in the absence of changes in the domestic credit component of the money supply any increase in the nominal demand for money must be equilibrated through the accumulation of foreign exchange reserves.

Empirical tests of the monetary approach seem to validate its main prediction: after compensation for real income growth differentials, those countries which expand domestic credit at a faster rate than the world average tend to run balance-of-payments deficits.[3] All of these studies are, however, of a long-run nature and follow the theoretical structure developed by Johnson which, among other things, assumes the constancy of all relative prices and an exogenous and known rate of real income growth. Short-run responses of the balance of payments to unexpected changes in relative prices or real income have not, to my knowledge, been the subject of empirical investigation within the framework of this approach. At a theoretical level, however, several papers[4] have concentrated on the dynamic path of the balance of payments following several types of disturbances within the framework of the monetary approach. Although not necessarily common to all of them, the most typical feature of the dynamic models is that, in the absence of domestic credit creation by the monetary authority, the balance-of-payments surplus equals the desired rate of accumula-

* Associate professor of economics, Columbia University. I want to thank George Borts and Rudiger Dornbusch for helpful comments. In a recent paper, Borts and James Hanson have independently arrived at results similar to those presented here.

[1] Among others, Robert Mundell, Jacob Frenkel (1971), Harry Johnson, Rudiger Dornbusch (1973a, b), and Frenkel and the author. See also the forthcoming collection of essays edited by Frenkel and Johnson.

[2] See, however, Frenkel (1975) for references of work on similar lines dating back to the middle 1700's.

[3] See, for example, several of the papers in Frenkel and Johnson.

[4] Dornbusch (1973a), Frenkel (1971), and Frenkel and the author among others.

tion of cash balances by individuals which in turn is taken to be a fraction of the difference between the desired and actual stocks. Thus, assuming that the exchange rate is equal to unity, the balance-of-payments surplus ($B_s$) becomes:

$$B_s = \gamma(M^d - M) \tag{1}$$

where $M^d$ and $M$ represent the desired and actual stocks of nominal cash balances and $\gamma$ is the constant speed of adjustment to the excess stock demand for money.

The desired stock of cash balances is in turn usually assumed to be proportional to the level of nominal income (thus ensuring a unit income elasticity of the demand for real cash balances):[5]

$$M^d = k \cdot Y$$

Nominal income is in turn defined as:

$$Y = \sum_i p_i q_i$$

where $p_i$ and $q_i$ are the nominal price and quantity produced of good $i$.

Assuming that the outputs of all goods produced at home are constant and also traded at fixed world prices,[6] the impact effect of an exogenous change in the world price of one good on the domestic balance of payments is

$$\partial B_s / \partial p_j = k\gamma q_j \geq 0$$

since at the instant of the price change the actual nominal stock of cash balances remains unchanged.[7] Thus, the prediction of this simple model is that an exogenous rise in any price generates a balance-of-payments surplus independently of whether the good in question is imported or exported, provided there is some domestic production of the good (otherwise there would be no effect at all). In particular, this simple version of the monetarist model predicts that a rise in the price of an importable will immediately generate (if at all) a balance-of-payments surplus. The explanation for the surplus is purely monetarist: the rise in the price increases nominal income and thus the demand for nominal cash balances; given the adjustment process formulated, individuals will immediately start accumulating cash balances, which brings about the surplus.

Needless to say, I find that particular result of the simple monetarist model unsatisfactory and in the next section the model is extended in order to allow for a rise in the price of an importable to generate a balance-of-payments surplus or deficit depending on the values of certain parameters, in particular, the volume traded of the good in question and the monetarist effect discussed above.

## II

In this section, I present an alternative formulation to the determination of the balance of payments which relies heavily on the specification of the role of money in the determination of real expenditures and is based on two well-known concepts, Milton Friedman's permanent income and J. M. Keynes' precautionary demand for money.

In Keynes' words, individuals hold money, among other reasons, to "... provide for contingencies requiring sudden expenditure ...". (p. 196). In my interpretation, should such a contingency occur, individuals will use their cash balances to soften the effect of this change on the time pattern of their *real* expenditures. As a proxy for that level of real expenditures which individuals will try to maintain in the short run, I take Friedman's concept of permanent income. This transitory use of cash balances should be independent of the behavior of the individual's long-run desired target money holdings which, also following Friedman, I assume to depend

[5] For a derivation of the above form of adjustment to monetary "disequilibrium" in the context of an intertemporal optimization model and the proof that that particular form *must* imply unit income elasticity of demand for money, see Dornbusch and Michael Mussa.

[6] I am thus ignoring the possibility of existence of nontraded goods in this simple version. An extensive analysis introducing nontraded goods into the monetarist model is done in Borts and Hanson. See also Dornbusch (1973b).

[7] Allowing for substitution in production will not affect this result since the substitution effect of a price change on nominal income is zero, provided that competition and full employment prevail, i.e., $\sum_i p_i \partial q_i / \partial p_j = 0$.

on the individual's permanent income. Should the unexpected event prove to be permanent and, say, reduce the permanent income of the individual, he will modify slowly his pattern of expenditures to accommodate his cash balances to the new long-run desired level; this may imply a further reduction or even accumulation of cash balances depending on whether the initial phase of dishoarding brought them above or below the new long-run desired level.

Algebraically, I postulate the following expenditure function.

$$e = y^p + \gamma(m - ky^p) \tag{2}$$

where all variables are defined in terms of a composite bundle of commodities to be defined below and

$e$ = real expenditures
$y^p$ = real permanent income
$ky^p$ = long-run desired real money holdings
$m$ = actual real money holdings
$\gamma$ = constant speed of adjustment to a stock excess demand for money

According to (2), real expenditures will exceed or fall short of permanent income only to the extent that individuals are not on their long-run demand for money. Although the model presented here could be extended to incorporate other assets alternative to money (possibly also performing the role of short-run shock absorbers), I will for simplicity assume that money is the only available asset and thus expenditures as defined in (2) represent the acquisition of goods by individuals. The term $\gamma(m-ky^p)$ in (2) is not, however, the actual accumulation of real cash balances. In an open economy with fixed exchange rates and no credit creation by the monetary authority, the actual change in the stock of money equals the balance-of-payment surplus ($b_s$), the difference between *current* real income and real expenditures:

$$\begin{aligned} dm/dt &= b_s = y - e \\ &= (y - y^p) - \gamma(m - ky^p) \end{aligned} \tag{3}$$

where $y$ stands for the current level of real income.

There are two main differences between the formulation of the balance of payments presented in (3) and that of the simple version presented in (1): (i) the term $(y-y^p)$ appears as a direct contribution to the balance-of-payment surplus; and (ii) the long-run demand for money depends on permanent rather than current income. Of course, when permanent and current income are equal, both expressions become identical.

The $(y-y^p)$ term in (3) characterizes the shock absorber nature of real cash balances in the short run since, to the extent that the term $\gamma(ky^p-m)$ represents *ex ante* desired accumulation of cash balances, it is clear that the desired rate will differ from the actual rate of accumulation by the full amount of the discrepancy between actual and permanent income.

### A. *Consumption and Production Technology*

Before going into the analysis of short- and long-run effects of price changes on the balance of payments we need to specify the units in which the real magnitudes in (2) are measured. This is particularly important in this model since one of the real variables, $y^p$, will be assumed to be unresponsive to any price change in the short run and thus the units in which it is measured become crucial. Obviously, the more important step here is the selection of the correct price level. In order to avoid problems which are unessential to the analysis and to provide an unambiguous definition of the price level it will be assumed that all goods are demanded in fixed proportions irrespective of income levels or relative prices. I also retain the assumption made in the last section that the outputs of all goods domestically produced are constant. These two assumptions will allow us to concentrate on macro-economic issues which will in any case be present in a more complete model where substitution effects are also accounted for.

The demand pattern is thus represented by:

$$C_i/C_j = a_i/a_j, \quad i, j: 1, \ldots, N, \text{ all } a \text{ constant}$$

and

$$\sum_i p_i C_i = E$$

where $C_i$ = demand for good $i$
$E$ = nominal expenditures
$p_i$ = nominal price of good $i$

Define the unit consumption basket as that basket that contains $a_i$ units of each good, or $(a_1, a_2, \ldots, a_N)$. The price of that basket is then

$$p = \sum_i p_i a_i$$

which will be our measure of the price level. It follows easily that

$$C_j = a_j \cdot E/p = a_j e$$

where $e$ stands for the number of unit baskets demanded (real expenditures).

On the production side, nominal income is

$$Y = \sum_i p_i q_i$$

and real income, $y$, is correspondingly defined as

$$y = Y/p = \sum_i p_i q_i / \sum_i p_i a_i$$

Real income responds to price changes according to

$$\partial y/\partial p_j = (q_j - a_j y)/p$$

and, if also $y=e$ which implies $C_j = a_j y$, it follows

$$\partial y/\partial p_j = (q_j - C_j)/p$$

which implies that a rise in the nominal price of good $j$ will increase real income if good $j$ is exported or decrease it if the good is imported.

Similarly, real cash balances $m$ are defined as

$$m = M/p$$

and the effect on real cash balances of a price change is given by:

$$\partial m/\partial p_j = -\, m a_j/p$$

### B. *Impact Effect of a Price Change on the Balance of Payments*

I will here maintain the assumption that the country is a price taker in all international markets and that there are no nontraded goods. The impact effect of a change in the price of good $j$ on the balance of payments can be readily obtained by differentiation of (3) holding $y^p$ constant. Starting with an initial situation where $b_s = 0$ and $y = y^p = e$, this derivative is

$$\text{(4)} \quad \partial b_s/\partial p_j = (1/p)[(q_j - C_j) + m a_j \gamma]$$

If the price increase corresponds to an exportable good $(q_j - C_j > 0)$, it is clear that the balance of payments must initially improve since both the increase in current real income (captured by the term $q_j - C_j$) and the reduction in actual real cash balances (captured by the term $m a_j \gamma$) will work towards a surplus. If however the price increase corresponds to an importable, the response of the balance of payments is ambiguous and more likely to be negative the larger the import volume of the good in question. Notice that the expression we have derived incorporates the standard monetarist effect of a price increase reducing real cash balances and thus contributing to a surplus. In addition it considers the full increase in the import bill (export bill) as a contribution to the balance-of-payments deficit (surplus) *independent of monetary considerations.* This latter effect, which has been disregarded so far in the short-run monetarist models, provides a plausible explanation for the observed fact of sharp changes in the prices of importables generating deficits in the balance of payments. However the model is flexible enough and allows for the pure monetarist effect to dominate when the amount traded of the good in question is small but the share of real expenditures on the good and the speed of adjustment to stock disequilibrium are large.

It should be clear from our analysis thus far that those short-run effects on the balance of payments of changes in the import or export bill are due exclusively to the short-run unresponsiveness of real expenditure to changes in the level of real income. As time goes on and individuals revise their permanent income level towards the new prevailing real income, expenditures will adapt and those effects will eventually disappear. This transitional period is the object of dynamic analysis contained in the next section.

### C. *Dynamic Effects of a Price Change on the Balance of Payments*

In deriving the impact effect of a price change I have assumed that permanent in-

come remained constant irrespective of the price change. It is obvious, however, that consumers will eventually have to adapt to this permanent price change and will thus have to modify their permanent income accordingly. How fast permanent income will be revised to the new level will depend on the degree with which individuals believe the price change is permanent and on the nature of the adjustment costs involved in changing the level of real expenditures. Remember we are using permanent income as a measure of that level of real expenditures households wish to maintain in the short run, presumably because there are adjustment costs involved in the process of changing the level of real expenditures. In what follows I will assume, for purposes of exposition only, that permanent income is adjusted through time according to the adaptive process

$$(5) \qquad dy^p/dt = b(y - y^p) \qquad b \text{ constant}$$

Also, from (3), real cash balances adjust through time according to

$$(6) \qquad dm/dt = (y - y^p) - \gamma(m - ky^p)$$

Given the actual value of real income, $y$, (5) and (6) describe the behavior over time of real cash balances (and by implication the balance of payments) and permanent income. The system (5)–(6) is linear and the matrix of the coefficients is

$$\begin{bmatrix} -b & 0 \\ -(1 - \gamma k) & -\gamma \end{bmatrix}$$

for which the trace is always negative and the determinant positive. Thus, the system (5)–(6) is globally stable, meaning that for any initial values for $m$ and $y^p$, both will eventually converge to the unique steady-state values $m^* = ky$ and $^*y^p = y$.

The final approach to the steady state must be asymptotic since the coefficient matrix of the dynamic system is triangular and thus the characteristic roots of the system are both real (and equal to $-b$ and $-\gamma$). The dynamic paths of $m$ and $y^p$ following a rise in the price of an importable are depicted in Figures 1a and 1b. There are two dynamic configurations depending on $(1-\gamma k)$ being

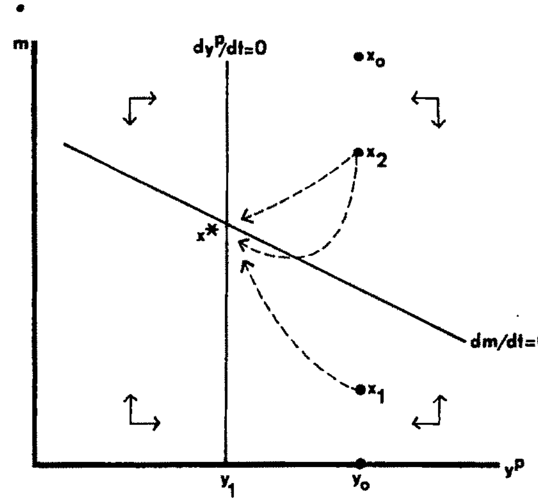



FIGURE 1a.

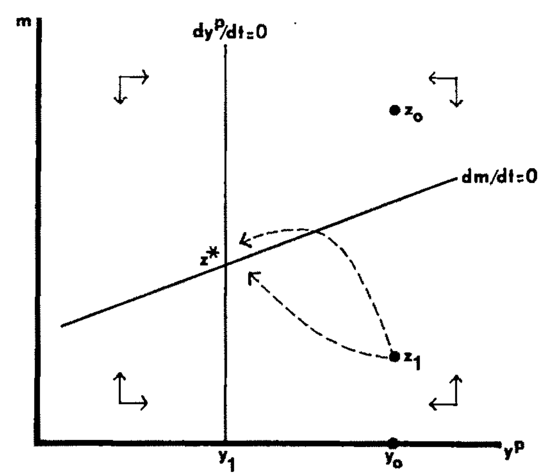



FIGURE 1b

larger or smaller than zero. Figure 1a shows the case where $(1-\gamma k)>0$ while Figure 1b illustrates the case where $(1-\gamma k)<0$. In both cases the schedule $dm/dt=0$ shows the pairs of $(m, y^p)$ for which the rate of change in the stock of real cash balances is zero according to (6). The slope of this schedule is:

$$dm/dy^p = -(1 - \gamma k)/\gamma$$

The vertical schedule $dy^p/dt=0$ shows that only for $y^p = y_1$ (the postprice change level of real income) will permanent income remain unchanged, according to the adaptive

process formulated in (5). The directions of the arrows show the movements of the variables in the plane and are derived from the signs of the coefficients in (5)–(6). Notice that Figure 1b, for which $(1-\gamma k)<0$, corresponds to the extreme situation where the monetary effects are the strongest, even in the short run; in this case the reader can verify from (2) that an increase in permanent income will actually be followed by a *fall* in real expenditures because the increased desired rate of hoarding will exceed the increase in permanent income. It is no surprise that in this case any price increase will generate a surplus when starting from a full equilibrium situation. This can be verified by substituting $ky$ for $m$ and $C_j$ for $a_jy$ in (4) after which the impact effect of a price change on the balance of payments becomes

$$(4') \quad \partial b_s/\partial p_j = (1/p)[q_j - C_j(1-\gamma k)]$$

which is always positive if $(1-\gamma k)<0$.

The initial positions before the rise in the price of the importable good are denoted by $x_0$ and $z_0$ in Figures 1a and 1b, respectively. At $x_0$ or $z_0$, real income was $y_0$ and, since it was a full-equilibrium position, it also was $y^p=y_0$ and $m=ky_0$. Following the price increase, both real income and real cash balances fall instantaneously. The new level of real income is denoted by $y_1$ in both figures. The change in $y$ also shifts the $dm/dt=0$ and $dy^p/dt=0$ schedules to the left such that at their new intersection (points $x^*$ or $z^*$), $m=ky_1$ and $y^p=y_1$. We can get an idea about the magnitude of the initial fall in real cash balances by noticing that the price rise may generate a surplus or a deficit in the case of Figure 1a and a surplus in Figure 1b. Since a surplus corresponds to an initial position where $dm/dt>0$, this corresponds to a fall in $m$ to points such as $x_1$ or $z_1$ in Figures 1a or 1b, respectively. If the price rise generates an instantaneous deficit in the balance of payments, the fall in $m$ must leave the system in an initial position where $dm/dt<0$, at a point like $x_2$ in Figure 1a. It is easy to see in the case of Figure 1a that if the initial impact effect is to bring the balance of payments into surplus, it will continue in surplus until the new steady state is reached (both $m$ and $y^p$ will remain in the region where $dm/dt>0$). If, however, the impact effect is to bring the balance of payments into deficit there are two possible outcomes: (i) it continues all the way into deficit until the new steady state is reached; or (ii) in later stages the deficit turns into a surplus which eventually becomes zero as the steady state is reached. In the case of Figure 1b, the initial surplus may be followed in later stages by a deficit.

An increase in the price of an exportable must be always followed by a balance-of-payments surplus although in later stages the surplus could turn into a deficit. This latter case could happen if expenditures do not adjust fast enough to the increased real income and thus more real cash balances accumulate than desired in the new steady state. The decumulation of those excess cash balances gives rise to the deficit in the later stages.

## REFERENCES


G. H. Borts and J. A. Hanson, "The Monetary Approach to the Balance of Payments," unpublished paper, Brown Univ. 1975.

R. Dornbusch, (1973a) "Currency Depreciation, Hoarding and Relative Prices," *J. Polit. Econ.*, July/Aug. 1973, *81*, 893–915.

———, (1973b) "Devaluation, Money and Nontraded Goods," *Amer. Econ. Rev.*, Dec. 1973, *63*, 871–80.

——— and M. Mussa, "Consumption, Real Balances and the Hoarding Function," *Int. Econ. Rev.*, June 1975, *16*, 415–21.

J. Frenkel, "A Theory of Money, Trade and the Balance of Payments in a Model of Accumulation," *J. Int. Econ.*, May 1971, *1*, 159–87.

———, "Adjustment Mechanisms and the Monetary Approach to the Balance of Payments: A Doctrinal Perspective," in E. Classen and P. Salin, eds., *Recent Issues in International Monetary Economics*, Amsterdam 1975.

——— and H. G. Johnson, *The Monetary Approach to the Balance of Payments*, London 1976.

——— and C. A. Rodriguez, "Portfolio Equilibrium and the Balance of Payments: A Monetary Approach," *Amer. Econ. Rev.*, Sept. 1975, *65*, 674–88.

M. Friedman, *A Theory of the Consumption Function*, Princeton 1957.

H. G. Johnson, "The Monetary Approach to

# The Distribution of State Incomes: Differential Growth of Sectoral Employment

By DONALD MITCHELL SMITH AND E. JAMES JENNINGS*

The degree of income concentration varies significantly among countries and among states of the United States. It is commonly accepted by economists that the concentration of income is less unequal in high income economies. The relation between low income concentration and high aggregate income is ascribed to such factors as the distribution of property income, the education of low income workers, and the relative size of the nonagricultural sector.

An implication of the hypothesized relation between income concentration and the relative size of nonagricultural employment is that intersectoral labor movements decrease income concentration. So far, this implication has not been tested. In this paper, the relation between changes in income concentration and changes in sectoral employment within states of the United States is tested directly.

The findings in this paper confirm the hypothesis that changes in income concentration are related to differential growth rates of sectoral employment within states. Between 1920 and 1950, and between 1950 and 1960, the change in employment of the agricultural vis-à-vis the nonagricultural sector is an important determinant of the change in income concentration. Between 1960 and 1970, the change in income concentration is related to the change in employment in the services sector vis-à-vis the combined agricultural and manufacturing sectors.

## I. Theory

Changes in the concentration of income may result from two major sources. The first source is increased wages such as that which result from an investment of human capital in otherwise low income workers. Ahmed Al-Samarrie and Herman Miller, p. 70, for example, find a strong negative correlation between Gini coefficients and median school years completed.

A second major source of changes in income concentration results from changes in the mix of employment in the economy. The historical decline of agricultural employment in the United States has increased the number of employees who work in manufacturing and services at levels of higher reported income. This hypothesis is weakly verified in a finding by Al-Samarrie and Miller that the Gini coefficient is positively related to labor earnings in agriculture relative to labor earnings in all industries.

Simon Kuznets offers some theoretical justification of the relation between income concentration and the proportion of nonagricultural employment. Kuznets views the aggregate distribution of income as derived from a combination of the income distributions in the rural (agricultural) and urban (manufacturing and service) populations. He stresses not only the difference in average per capita incomes, but also a potential difference in the income distribution between the two sectors (pp. 7, 8). A shift of employment from the agricultural to the nonagricultural sector, argues Kuznets, will increase income inequality because the urban sector exhibits a greater income inequality, and because employees who are left in the agricultural sector continue to receive relatively low incomes.

If the income inequality is not greatly different between the two sectors, however, a shift from a largely agrarian society to an industrial and services society will first cause an increase, and then an eventual decrease,

* Economist, Office of Economic Impact Analysis, Federal Energy Administration, and doctoral candidate, Southern Methodist University, respectively. The ideas in the paper are those of the authors and do not necessarily represent opinions or policies of the federal government.

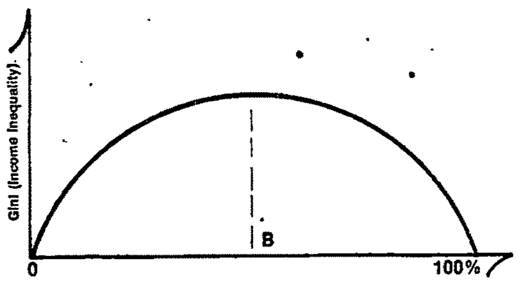



FIGURE 1. PERCENTAGE OF EMPLOYMENT IN THE NONAGRICULTURAL SECTOR

(As the percentage of nonagricultural employment increases from 0 to 100, the degree of income inequality first increases, then decreases)

in income inequality. For the sake of argument, assume complete income equality within each sector, but that income per worker in sector $A$ (agricultural) is less than that in sector $NA$ (nonagricultural). If the economy begins with almost all employment in sector $A$, the income distribution will be equal (the Gini almost zero). As workers transfer to the higher wage $NA$ sector, the distribution of income becomes more unequal (the Gini increases). With continued transfer to the $NA$ sector, most of the employment will be at the high wage level, and the distribution of income will become less unequal. In Figure 1, the Gini coefficient is shown to increase and then decrease as the proportion of nonagricultural employment increases.

Al-Samarrie and Miller's negative empirical relation between the Gini and the proportion of total earnings in the nonagricultural sector is probably due to the fact that most states of the United States were beyond point $B$ in 1950 and 1960, when their cross-section data was collected.

This paper studies only the concentration of income in relation to the relative sizes of employment within sectors of a state's economy. It is assumed that all workers in each sector of a state receive the same wage, although the wage level differs between sectors. Therefore, the distribution of income (measured by wage income per worker) is equal within each sector.[1] The unequal distribution of income within a state (composed of two sectors) is due to the mix of employment and the wage differential between the sectors.

Changes in income concentration over periods when the wage differential between the sectors is approximately constant will be caused by the varying growth rates of the two sectors. The causes of varying growth rates of sectoral employment have been treated at length in the economic literature, and will not be incorporated into this analysis. However, it is implicitly assumed that labor moves between states and between sectors of states in response to wage differentials and other indicators of the prospects for higher income.[2]

## II. Gini Coefficient and Variance of Incomes

Since the focus of this paper is on the relation between the concentration of incomes and the sectoral mix of employment, the concentration of income will be measured from Internal Revenue Service ($IRS$) data on personal incomes. An advantage of the $IRS$ data compared to the U.S. Census of Population data is that transfer payments are excluded. These data are available from 1920, and may be used in an analysis of income concentration of states.

Because the change in the Gini coefficient may be positively or negatively related to a relative increase in nonagricultural employment the analysis will be made through the intervening calculation of the variance of income, measured between two sectors each having a constant but different wage level. The use of the variance allows direct inclusion of the relative size of the two sectors, and will allow mathematical adjustment of the alternating positive and later negative relation between the Gini and the proportion of nonagricultural employment.

The degree of income concentration is conventionally measured by Gini's coefficient of concentration, defined by

[1] Kuznets finds that the difference in income distribution between the agricultural and nonagricultural sectors "in the years after the second world war seems to have disappeared" (p. 8).

[2] Smith, for example, finds such a model consistent with the growth experience of states and sectors of states in the United States. However, Lloyd Ulman finds that labor migration does not induce substantial reductions of wage differentials.

$$(1)\qquad G = \Delta_1/2u_1'$$

where

$$(2)\qquad \Delta_1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} | x - y |\, dF(x)dF(y)$$

and $x$ and $y$ are cumulative percentages of income received by each income group, and $u_1'$ is the mean. Following M. G. Kendall and A. Stuart (p. 47), we define

$$(3)\qquad E^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - y)^2 dF(x)dF(y)$$

$$(4)\qquad E^2 = 2u_2$$

where $u_2$ is the variance.

On inspection we note that the difference between $\Delta_1$ and $E^2$ is that one calculation involves $|x-y|$, whereas the other involves $(x-y)^2$. Since $|x-y|$ is a monotonic transformation of $(x-y)^2$, $\Delta_1$ is a monotonic transformation of $E^2$. Furthermore, since $G$ is simply $\Delta_1$ divided by a positive constant, $G$ must be a monotonic transformation of $E^2$, which is equal to twice the variance. Thus $G$ is also a monotonic transformation of the variance.

By the assumption of a single wage in each sector of a state, the variance of a state's income is

$$(5)\qquad V = \frac{L_i(W_i - \overline{W})^2 + L_j(W_j - \overline{W})^2}{L_n}$$

where, in each state,

$\overline{W}$ = mean wage
$W_i$ = wage in sector $i$
$W_j$ = wage in sector $j$
$L_i$ = employment in sector $i$
$L_j$ = employment in sector $j$
$L_n$ = total state employment

It is assumed that wage levels are constant throughout the period under examination. Thus, only the mean wage changes with changes in employment.

The change in the variance is

$$(6)\qquad dV = \frac{L_j}{L_n^2}(W_i - W_j)^2 dL_i + \frac{L_i}{L_n^2}(W_i - W_j)^2 dL_j - 2\frac{L_iL_j}{L_n^3}(W_i - W_j)^2 dL_n$$

There are three components of change in the variance which are variables in this model: $dL_i$ is the change of employment in the declining sector; $dL_j$ is the change in employment in the growing sector; and $dL_n$ is the change in employment in the state which is net migration from another state, and is distributed among the two sectors in the same proportion as the existing mix of employment.

In equation (6) it is clearly seen that an increase in $dL_j$ which equals a decrease in $dL_i$ will bring about a positive or negative change in $dV$, depending on the relative sizes of the coefficients of $dL_i$ and $dL_j$.

In a primarily agrarian economy a positive $dL_i$ equal to negative $dL_j$ will increase $V$, due to $(W_i-\overline{W})>(W_j-\overline{W})$. Therefore, a $dL_j=-dL_i$ will cause a decline in $V$ (or the Gini coefficient). In general, however, interstate migration will affect $dL_i$ and $dL_j$ differentially so that $dL_i=dL_j$ will not be found empirically.

Because $V$ is a monotonic transformation of $G$ we substitute $dG$ for $dV$ in equation (6). The coefficients of $dL_i$, $dL_j$, and $dL_n$ are included as variables in the equation which tests the relation between changes in the Gini and changes in sectoral employment.

$$(7)\qquad dG = \alpha_0 + \alpha_1X_1 + \alpha_2X_2 + \alpha_3X_3 + e$$

where

$$X_1 = (W_{ia} - W_{ja})^2\frac{L_{jb}}{L_{nb}^2}[L_{ic} - L_{ib}]$$

$$X_2 = (W_{ia} - W_{ja})^2\frac{L_{ib}}{L_{nb}^2}[L_{jc} - L_{jb}]$$

$$X_3 = (W_{ia} - W_{ja})^2 2\frac{L_{ib}L_{jb}}{(L_{nb})^3}[L_{nc} - L_{nb}]$$

and $e$ is a vector of randomly distributed error terms having a mean of zero.

The second subscript $b$ is the first year of the time period, $c$ is the last year of the time period, and $a$ is the average of the first and last year's data for the period. For the 1920–50 and 1950–60 periods, the agricultural ($i$) and nonagricultural ($j$) sectors are used to

compute the independent variables. For the 1960–70 period, however, the data are arranged such that the growing sector ($j$) is composed of services, and the declining sector ($i$) is an aggregate of the manufacturing and agricultural sectors.[3]

The arrangement of the data for the first two periods is consistent with the literature pertaining to the historical decline of agriculture in the United States. However, by 1960 the migration from the agricultural sector had diminished considerably. For this reason the growing services sector was isolated from manufacturing. The agricultural sector was combined with manufacturing for empirical practicality.[4]

By inspection of (6) and (7) $\alpha_1$ and $\alpha_2$ are greater than zero and $\alpha_3$ is negative. If $dG$ equalled $dV$ (rather than being a monotonic transformation), $\alpha_1$ and $\alpha_2$ would equal plus one, and $\alpha_3$ would equal minus one. Thus, the null hypothesis is that $\alpha_1$, $\alpha_2 \leq 0$, and $\alpha_3 \geq 0$. Rejection of this null hypothesis will allow acceptance of the alternate hypothesis that $\alpha_1$, $\alpha_2 > 0$; $\alpha_3 < 0$ which will verify the theory.

### III. Empirical Results

In Table 1 the least squares regression coefficients are presented. All coefficients except one are significantly greater than zero and of the expected sign. Thus the model is found to be in general agreement with the hypothesis that the change in the distribution of income is related to changes in sectoral employment.

From equation (6) it is seen that the variables $X_1$ and $X_2$ refer to weighted changes in sectoral employment, with net change in state employment $L_n$ held constant. The variable $X_3$ picks up the change in $G$ due to increased state labor force, with the intersectoral movement held constant.

The coefficient of $X_2$ is not significantly different from zero for the 1960–70 period. This result may be due to the combination of manufacturing and agricultural sectors when one is growing and the other declining simultaneously in some states.

The interpretation of the regression results may be assisted by an example using a single state for the 1920–50 period. The data were arranged so that the Gini coefficient could vary from zero to 1,000. In Alabama, migration from the agricultural sector of 298,900 workers is associated with a decline in the Gini of $-200.2$, when the weight (obtained from Table 1) is multiplied by the coefficient estimate. Similarly, the migration of 425,300 workers to the combined manufacturing and services sector is associated with an increase in the Gini of 231.4.

The relatively low $R^2$ in all equations requires some explanation. In the first place, the income data used in generating the independent variables are based upon only one of the sources of income, namely, labor. Because the proportion of property income varies between states, it may differentially impact state Gini coefficients. Therefore, a portion of variance in state Gini coefficients lies outside the explanatory power of the independent variables. Secondly, there exists a variance of income within sectors which will be incorporated in the Gini coefficient but is kept out of reach of this study by data restrictions. Thirdly, a model using many sectors rather than only two would have captured much more intrasectoral income variance. A fourth factor causing the relatively low $R^2$ may be due to the assumption that wage levels are constant for each sector during the period of analysis. Changing levels of employment may change capital-labor ratios and thus change wage levels. A decrease in wage level in the grow-

[3] A three-sector analysis was found to be impractical for statistical analysis because of the complexity of taking the total differential of the variance when three sectors was included. A regression was run on the agricultural vs. nonagricultural sector for the 1960–70 period. However, this yielded nonsignificant coefficients for $X_1$ and $X_2$.

[4] The data used in calculating the independent variables for 1920 and 1950 came from E. S. Lee et al. In the 1960 and 1970 periods nonagricultural employment and agricultural employment came from the *U.S. Census of Population, General Social and Economic Characteristics*. Manufacturing employment is from the *U.S. Census of Manufactures*, 1960 and 1970. Agricultural and nonagricultural income is from the *Survey of Current Business*, 1960 and 1970. The generation of the Gini coefficients is in accordance with James Morgan. The raw income data are from *Personal Income by States* 1920, 1950, 1960, and 1970, published by the *IRS*.

TABLE 1—REGRESSION COEFFICIENTS[a]

| Independent Variables | Dependent Variables | | |
|---|---|---|---|
| | Agricultural vs. Nonagricultural | | Service vs. Manufacturing and Agriculture |
| | $dG$ 1920–50 | $dG$ 1950–60 | $dG$ 1960–70 |
| $C$ | 80.4 | 25.4 | 13.6 |
| | (7.66)[b] | (9.00)[b] | (4.41)[b] |
| $X_1$ | $.179 \times 10^{-2}$ | $.780 \times 10^{-4}$ | $.22 \times 10^{-4}$ |
| | (2.32)[b] | (1.71)[b] | (1.91)[b] |
| $X_2$ | $.120 \times 10^{-2}$ | $.68 \times 10^{-3}$ | .64 |
| | (2.42)[b] | (2.58)[b] | (.29) |
| $X_3$ | $-.160 \times 10^{-2}$ | $-.700 \times 10^{-3}$ | −2.9 |
| | (1.88)[b] | (2.26)[b] | (2.14)[b] |
| $R^2$ | .16 | .14 | .13 |

[a] $t$-values are in parentheses.
[b] Significantly greater (or less than) zero at 5 percent.

ing sector would, for example, tend to decrease the unequal distribution of income.

## IV. Conclusion

This paper focuses on one component of the many influences on the concentration of income. Specifically, changes in the sectoral mix of employment are related to changes in the variance of state incomes which are demonstrated to be related directly to changes in Gini's coefficient of income concentration. An increase of employment in the high wage sector may cause an increase or decrease in the variance, depending on the relative size of income payments in the sector. For purposes of statistical testing, changes in sectoral employment are appropriately weighted and then regressed on the change in Gini coefficient.

This simple model, which omits other important influences on income concentration such as property income and the great diversity of incomes within sectors, explains about 10 to 20 percent of the variation of state Gini coefficients.

These results tend to confirm Kuznets' hypothesis that the process of economic growth (involving a shift from an agrarian to manufacturing and service economy) influences the concentration of income. In advanced stages of growth, however, the relatively faster growth rate of services vis-à-vis manufacturing is an important influence on the concentration of income.

Further research in the direction taken in this paper would incorporate the covariation of wage levels with changes in employment. The changing wage levels would then possibly influence the change in income concentration resulting from intersectoral migration.

## REFERENCES

D. J. Aigner and A. J. Heins, "On the Determinants of Income Inequality," *Amer. Econ. Rev.*, Mar. 1967, *57*, 175–81.

A. Al-Samarrie and H. P. Miller, "State Differentials and Income Concentration," *Amer. Econ. Rev.*, Mar. 1967, *57*, 59–72.

M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics*, Vol. 1, London 1969.

S. Kuznets, "Economic Growth and Income Inequality," *Amer. Econ. Rev.*, Mar. 1955, *45*, 1–28.

E. S. Lee et al., *Population Redistribution and Economic Growth, United States, 1870–1950*, Philadelphia 1957.

J. Morgan, "The Anatomy of Income Distribution," *Rev. Econ. Statist.*, Aug. 1962, *44*, 270–83.

D. M. Smith, "Regional Growth: Interstate and Intersectoral Factor Reallocation," *Rev. Econ. Statist.*, Aug. 1974, *56*, 353–59.

——— and J. E. Jennings, "Economic Growth and the Distribution of Income," working

pap. no. 45, Southern Methodist Univ.

L. Ulman, "Labor Mobility and the Industrial Wage Structure in the Postwar United States," *Quart. J. Econ.*, Feb. 1957, *79*, 73–97.

**Internal Revenue Service**, (*IRS*) *Personal Income by States*, 1920, 1950, 1960, 1970, Washington.

**U.S. Bureau of the Census**, *U.S. Census of Population:* 1960, 1970; *General Social and Economic Characteristics*, Washington 1961, 1971.

**U.S. Office of Business Economics**, *Surv. Curr. Bus.*, Washington 1960, 1970.

# Multiperiod Consumption-Investment Decisions: A Correction

By EUGENE F. FAMA*

My 1970 paper attempts to provide a multiperiod setting for hypotheses about investor behavior derived from one-period portfolio models. The main result of the paper is that if the investor is risk averse (his Von Neumann-Morgenstern utility function for lifetime consumption is strictly concave) and markets for consumption goods and portfolio assets are perfect, then the investor's behavior in the market in any period is indistinguishable from that of a risk-averse expected-utility maximizer who has a one-period horizon. William Ziemba has recently pointed out, however, that the proof of the theorem on which this conclusion is based is incomplete. I hope now to give a complete proof.

In the framework of my earlier paper, the consumption-investment decision facing the investor at any time t can be expressed in recursive form as

$$(1)\quad U_t(C_{t-1}, w_t \mid \beta_t) = \max_{c_t, H_t} \int_{\beta_{t+1}} U_{t+1}(C_t, H_t R(\beta_{t+1})' \mid \beta_{t+1}) dF_{\beta t}(\beta_{t+1})$$

$$\text{subject to} \quad 0 \leq c_t \leq w_t, \qquad H_t i' = w_t - c_t$$

Briefly, $H_t$ is the (row) vector of dollars invested in individual securities at time t; $R(\beta_{t+1})$ is the vector of returns realized on individual securities at time t+1 when the state of the world at t+1 is $\beta_{t+1}$; $F_{\beta t}(\beta_{t+1})$ is the distribution function for $\beta_{t+1}$ conditional on $\beta_t$; $c_t$ and $w_t$ are consumption and wealth at time t; $C_t$ is the vector of lifetime consumptions to time t; $i$ is the sum vector; and

$$(2)\qquad w_{t+1} = H_t R(\beta_{t+1})'$$

Finally, for any t, the function $U_t(C_{t-1}, w_t \mid \beta_t)$ shows the maximum expected utility of lifetime consumption if the consumer is in state $\beta_t$ at time t, his wealth is $w_t$, his past consumption was $C_{t-1}$, and optimal consumption-investment decisions are made at time t and at all future decision times.

As in most dynamic programming models, expression (1) represents the decision problem for any time t in terms of a derived objective function, $U_{t+1}(C_t, w_{t+1} \mid \beta_{t+1})$, which is explicitly a function of variables for t+1 and earlier periods, but which in fact summarizes the results of optimal decisions at t+1 and subsequent periods for all possible future events. Thus the recursive relation (1) represents the multiperiod problem as a sequence of "one-period" problems, but at any stage in the process the objective function used to solve the one-period problem summarizes optimal decisions for all future periods. The major result of the 1970 paper is that if the consumer's utility function for lifetime consumption has properties characteristic of risk aversion, then for all t the derived functions $U_t(C_{t-1}, w_t \mid \beta_t)$ also have these properties. From this it follows that each period the consumer behaves as if he were a risk-averse expected-utility maximizer with a one-period horizon. The formal basis for this result is the following:

PROPOSITION: *If* $U_{t+1}(C_t, w_{t+1} \mid \beta_{t+1})$ *is strictly increasing and strictly concave in* $(C_t, w_{t+1})$, *then* $U_t(C_{t-1}, w_t \mid \beta_t)$ *is strictly increasing and strictly concave in* $(C_{t-1}, w_t)$.

PROOF:

Let $c_t^*$, $H_t^*$ and $\tilde{c}_t$, $\tilde{H}_t$ be the optimal values of $c_t$ and $H_t$ in (1) for two vectors $(C_{t-1}^*\ w_t^*)$ and $(\tilde{C}_{t-1}, \tilde{w}_t)$ that differ in at least one element. Let

$$(3)\quad \begin{aligned} \hat{C}_{t-1} &= \alpha C_{t-1}^* + (1-\alpha)\tilde{C}_{t-1} \\ \hat{w}_t &= \alpha w_t^* + (1-\alpha)\tilde{w}_t \end{aligned}$$

* Professor of finance, Graduate School of Business, University of Chicago. I am grateful for the comments of Gary Eppen and Nicholas Gonedes, and for the financial support of the National Science Foundation.

$$\hat{c}_t = \alpha c_t^* + (1-\alpha)\tilde{c}_t$$
$$\hat{H}_t = \alpha H_t^* + (1-\alpha)\tilde{H}_t$$
$$\hat{w}_{t+1} = \alpha w_{t+1}^* + (1-\alpha)\tilde{w}_{t+1} = \alpha H_t^* R(\beta_{t+1})' + (1-\alpha)\tilde{H}_t R(\beta_{t+1})'$$

The strict concavity of $U_{t+1}$ in $(C_t, w_{t+1})$ implies that for $0<\alpha<1$,

$$(4)\quad U_{t+1}(\hat{C}_{t-1}, \hat{c}_t, \hat{H}_t R(\beta_{t+1})' \mid \beta_{t+1}) \geq \alpha U_{t+1}(C_{t-1}^*, c_t^*, H_t^* R(\beta_{t+1})' \mid \beta_{t+1}) + (1-\alpha) U_{t+1}(\tilde{C}_{t-1}, \hat{c}_t, \tilde{H}_t R(\beta_{t+1})' \mid \beta_{t+1})$$

This expression can only hold as an equality when

$$(5)\quad (C_{t-1}^*, c_t^*, w_{t+1}^*) = (\tilde{C}_{t-1}, \tilde{c}_t, \tilde{w}_{t+1})$$

Since $(C_{t-1}^*, w_t^*)$ and $(\tilde{C}_{t-1}, \tilde{w}_t)$ are assumed to differ in at least one element, (5) can only hold when the optimal allocations for two *unequal* wealth levels $w_t^*$ and $\tilde{w}_t$ imply $c_t^* = \tilde{c}_t$. Moreover, although $w_t^* \neq \tilde{w}_t$ and $c_t^* = \tilde{c}_t$ necessarily imply $H_t^* \neq \tilde{H}_t$, if (5) is to hold we must have

$$(6)\quad w_{t+1}^* = H_t^* R(\beta_{t+1})' = \tilde{H}_t R(\beta_{t+1})' = \tilde{w}_{t+1}$$

That is, in spite of unequal allocation vectors $H_t^*$ and $\tilde{H}_t$, the return vector $R(\beta_{t+1})$ leads to equal values of $w_{t+1}^*$ and $\tilde{w}_{t+1}$.

It is reasonable to assume, however, that the joint distribution of security returns is nonsingular so that there are necessarily possible values of $R(\beta_{t+1})$ for which unequal $H_t^*$ and $\tilde{H}_t$ imply unequal $w_{t-1}^*$ and $\tilde{w}_{t+1}$, in which case (4) holds as strict inequality. As long as there is one possible value of $\beta_{t+1}$ for which (4) must hold as a strict inequality, we have

$$(7)\quad \int_{\beta_{t+1}} U_{t+1}(\hat{C}_{t-1}, \hat{c}_t, \hat{H}_t R(\beta_{t+1})' \mid \beta_{t+1}) \cdot dF_{\beta t}(\beta_{t+1})$$
$$> \alpha \int_{\beta_{t+1}} U_{t+1}(C_{t-1}^*, c_t^*, H_t^* R(\beta_{t+1})' \mid \beta_{t+1}) \cdot dF_{\beta t}(\beta_{t+1}) + (1-\alpha)\int_{\beta_{t+1}} U_{t+1}(\tilde{C}_{t-1}, \tilde{c}_t, \tilde{H}_t R(\beta_{t+1})' \mid \beta_{t+1}) dF_{\beta t}(\beta_{t+1})$$
$$= \alpha U_t(C_{t-1}^*, w_t^* \mid \beta_t) + (1-\alpha) U_t(\tilde{C}_{t-1}, \tilde{w}_t \mid \beta_t)$$

Since the consumption-investment decision implied by $\hat{c}_t$, $\hat{H}_t$ is not necessarily an optimal allocation of $\hat{w}_t$

$$(8)\quad U_t(\hat{C}_{t-1}, \hat{w}_t \mid \beta_t) \geq \int_{\beta_{t+1}} U_t(\hat{C}_{t-1}, \hat{c}_t, \hat{H}_t R(\beta_{t+1})' \mid \beta_{t+1}) dF_{\beta t}(\beta_{t+1})$$

so that

$$(9)\quad U_t(\hat{C}_{t-1}, \hat{w}_t \mid \beta_t) > \alpha U_t(C_{t-1}^*, w_t^* \mid \beta_t) + (1-\alpha) U_t(\tilde{C}_{t-1}, \tilde{w}_t \mid \beta_t)$$

Thus $U_t$ is strictly concave in $(C_{t-1}, w_t)$. The monotonicity of $U_t$ in $(C_{t-1}, w_t)$ follows easily from the monotonicity of $U_{t+1}$ in $(C_t, w_{t+1})$, and the proposition is established.

## REFERENCES


E. F. Fama, "Multiperiod Consumption-Investment Decisions," *Amer. Econ. Rev.*, Mar. 1970, *60*, 163–74.

W. T. Ziemba, "The Behavior of a Firm Subject to Stochastic Regulatory Review: Comment," *Bell. J. Econ.*, Autumn 1974, *5*, 710–12.

# IN MEMORIAM

## CALVIN BRYCE HOOVER

## 1897-1974

Calvin Bryce Hoover, James B. Duke Professor Emeritus of Economics at Duke University, died June 24, 1974, in his seventy-seventh year. Born in Berwick, Illinois, Hoover received an A.B. degree from Monmouth College in 1922 and the Ph.D. from the University of Wisconsin in 1925. That same year he went to Duke as an assistant professor in the department which remained his academic home until his retirement and death.

Among both graduate and undergraduate students he was known as an outstanding teacher. For many years in the advanced theory courses his students learned to apply previously studied economic principles both to micro and macro problems. His most popular courses, however, were Economic Systems and Economic Functions of the State, to which he brought his extensive experiences with communism and fascism, and with numerous aspects of federal bureaucracy. His academic life included twenty years service as chairman of the department of economics and for ten of those years, as Dean of Duke's Graduate School of Arts and Sciences.

A reading of his autobiography, "Memoirs of Capitalism, Communism, and Nazism," provides understanding of Hoover's insight into the concerns of both the small farmer and the less skilled worker. His own boyhood experiences probably sharpened his interest in the plight of the southern farmer and worker and led to his aiding in the organization of the Committee of the South and to serving for many years as that committee's Director of Research. With Professor Ben Ratchford, he published in 1951 a penetrating monograph, *Economic Resources and Economic Policies of the South.*

Hoover's chief contributions to economic literature were his analyses of the development and operation of Russian communism and German nazism. His first book, *Economic Life in Soviet Russia*, was written after a year spent in Russia during the great Stalinist purges. He was in Germany in 1932–33 during Hitler's rise to power and along with Lord Keynes (a long-time friend of Hoover's) saw that Hitler was using his rearmament program as a weapon to restore full employment and that an underemployed economy, when stimulated, may produce both guns and butter. From his German experiences came his second book, *Germany Enters the Third Reich.*

Having seen totalitarianism at work, Calvin Hoover became an ardent, but never uncritical, believer in more liberal systems. His many books and articles dealing with economic theory, nontotalitarian economies, and with state economic policies were well-reasoned economic analyses, though never to the neglect of critical political factors when pertinent. His writings and numerous addresses always reflected the logic and reasoning of the trained economist.

Hoover's many nonacademic assignments in government in both peace time and wartime drew heavily upon his skills as advisor and administrator. He was proud of his services as a private and noncommissioned officer during the First World War where, according to him, he learned nonacademic language which he could use effectively at the proper time and place.

During the New Deal days he was an economic advisor and administrator under Henry Wallace in the Department of Agriculture. When the United States entered the war, Hoover, because of his experiences in Nazi Germany and the Soviet Union, was recruited by General Donovan for the intelligence service later called O.S.S. He played various roles while serving, eventually becoming head of Northern European operations based in Stockholm. According to one competent observer, "His network was considered the hottest and most professional of all United States' intelligence operation." For his wartime activities Hoover was

awarded the Medal of Freedom by President Truman.

After the war, in 1945, he returned to Berlin where under Generals Clay and Draper, he prepared the Hoover Plan for restoring the German economy, a program at variance with the earlier Morgenthau line to prevent German economic recovery. Later Hoover was one of the principal initiators and developers of the Marshall Plan, which he helped administer as special advisor to the *U.S.* representative in Europe for the Economic Cooperation Administration.

His Duke friends and colleagues called him Bryce and no memorial statement would be complete without some mention of Hoover as a colleague. He could be firm, perhaps even stubborn, but he seldom gave advice without a preface that advice that had to be taken was not worth getting. His enormous fund of always interesting and appropriate stories, together with his warm personality, made him a great companion whether he was discussing economic problems or recounting some of the events of an active life or simply telling stories. And he seldom if ever repeated himself.

The honorary degrees he received from Columbia, Monmouth College, Duke University and Case-Western Reserve and his election to the presidency of the Southern Economic Association, of The Association for Comparative Economies, and of the American Economic Association, showed recognition of an able economist and a devoted public servant.

# NOTES

The eighty-ninth annual meeting of the American Economic Association will be held in Atlantic City, New Jersey, September 16–18, 1976. The Employment Registry and Center will be open from September 15 to 18.

As the annual meeting is being held at an early point in the academic year, it may not be possible to complete all job placement activity through the offices of the Employment Registry and Center. Accordingly, the American Economic Association will hold a special job placement meeting for two days at the O'Hare Hilton in Chicago's O'Hare International Airport, January 7–9, 1977.

---

*Annual Meeting Placement Service*

The Placement Service at the 1976 annual meetings of the Allied Social Science Associations in Atlantic City will begin operation on September 15, the day before sessions begin. Applicants and employers will be able to attend more sessions with a day set aside entirely for labor market transactions. This service will be located at the Convention Placement Center in the Hotel Shelbourne. It will be open from 10:00 A.M. to 5:00 P.M., September 15; 8:30 A.M. to 5:00 P.M., September 16–17; and 8:30 A.M. to 12:00 noon September 18.

---

Economists who are *strongly* oriented toward the humanities, who use humanistic methods in their research, and who will be participating in meetings abroad that are concerned with the humanistic aspects of their discipline are eligible to apply for small travel grants of the American Council of Learned Societies. Grants are likely to cover only lowest cost excursion fares and will rarely exceed 50 percent of full economy-class fares. Specifically, economists may be eligible if (a) they deal with the history of economic thought or economic history, and (b) if their approach is qualitative and descriptive rather than quantitative and statistical. Conferences dealing with the establishment of social policy or legislation are ineligible. The deadlines for applications to be received in the office of the American Economic Association are: meetings scheduled between July and October, March 1; for meetings scheduled between November and February, July 1; for meetings scheduled between March and June, November 1. Application forms may be obtained from C. Elton Hinshaw, Secretary, American Economic Association, 1313 21st Avenue South, Nashville, Tennessee 37212.

---

The History of Economics Society will hold its next annual conference at the University of California, Riverside, on March 24–26, 1977. Persons wishing to present papers should submit an abstract in duplicate together with a separate sheet containing author's name, address, professional affiliation, position, and telephone number; title of paper; and whether or not the author is a member of the Society. The deadline is November 15, 1976. Persons wishing to serve as discussants are invited to submit full information including major areas of interest within the history of economics. All communications regarding the conference should be addressed to President-elect Carl G. Uhr, History of Economics Society, Department of Economics, University of California, Riverside, CA 92502.

---

The Eastern Community College Social Science Association will hold its 1976–77 conference at the Webster Hall Hotel in Pittsburgh, Pennsylvania, April 14–16, 1977. Those interested in presenting papers, participating in panel discussions, or presenting media demonstrations should send an abstract to George H. Skau, Program Chairman, Social Science Department, Bergen Community College, 400 Paramus Road, Paramus, NJ 07562. The deadline is September 30, 1976.

---

The African Studies Program, University of Wisconsin, announces a new journal, *African Economic History*, to be published twice a year. It will focus on recent economic change in Africa as well as on the colonial and precolonial economic history of the continent. Subscriptions are $4.50 to individuals and $7.00 to institutions and organizations outside of Africa ($4.50 for those in Africa), to be remitted in U.S. currency and made payable to the University of Wisconsin. Manuscripts and subscriptions should be sent to The Editors, *African Economic History*, 346B Agricultural Hall, 1450 Linden Drive, Madison, Wisconsin 53706.

---

The Southern Regional Science Association will hold its 1977 meetings in Birmingham, Alabama, on April 14–15, 1977. Scholars interested in presenting papers on any subject involving analysis of the spatial dimensions of human activity are invited to submit one-page abstracts by October 1, 1976 to the program chairman, Professor Shirley F. Weiss, Center of Urban and Regional Studies, University of North Carolina, Chapel Hill NC 27514.

---

*Meetings of Regional Economic Associations*

Eastern Economic Association Meeting, April 14–16, 1977, Hartford, CT.

Midwest Economic Association Meeting, March 31–April 2, 1977, St. Louis, MO.

Western Economic Association Meeting, June 19–23, 1977, San Francisco, CA.

Missouri Valley Economic Association Meeting, February 24–26, 1977, St. Louis, MO.

---

Omicron Delta Epsilon, the International Honor Society in Economics, invites the submission of entries for the ninth year of the Irving Fisher Graduate Monograph and Frank W. Taussig Award Undergraduate Competitions. The Fisher Award consists of $1,000 and publication as a book by Princeton University Press, subject to approval of its editorial board. In addition, the winner will be invited to submit a paper based on the winning entry to the *American Economic Review*. The recommendations of the Final Selection Board of the Competition will be considered by the *Review* in the refereeing process. All finalists will be invited to submit a paper for publication in *The American Economist*. The Taussig Award consists of $100 and publication in *The American Economist*. Entries for the Fisher Award should be submitted to Departmental Selection Committees by January 1, 1977 and entries for the Taussig Award by May 15, 1977. They will be judged by the Selection Board consisting of Professors Frank H. Hahn, Dale W. Jorgenson, Robert M. Solow, Arnold Hahn, Dale W. Jorgenson, Robert M. Solow, Arnold Zellner, and Egon Neuberger (editor). Anyone interested in entering the competition should contact Professor E. Neuberger, Editor, Economic Research Bureau, State University of New York, Stony Brook, NY 11794, for information on entrance requirements.

---

Under a joint program of the Social Science Research Council and the American Council of Learned Societies, fellowships are offered for nine to eighteen months of dissertation research on contemporary Western European affairs. Applications are invited from students in all social sciences and humanities, but particular attention is given to disciplines in which relatively less attention has been devoted to Western Europe, such as anthropology, economics, social psychology, and sociology. Requests for further information and application forms should be sent to Social Science Research Council, Western European Program, 605 Third Avenue, New York, NY 10016.

---

## *Deaths*

William C. Bagley, Jr., associate professor, department of economics, Rutgers–The State University, Feb. 29, 1976.

Earl Crockett, professor of economics, Brigham Young University, Dec. 2, 1975.

Thomas A. Martinsek, professor of economics, Southern Illinois University, Aug. 15, 1975.

Sidney E. Rolfe, professor of finance, Long Island University, and research associate, Center of International Studies, Massachusetts Institute of Technology, Mar. 10, 1976.

Emerson P. Schmidt, former director of economics research, Chamber of Commerce of the United States, Aug. 8, 1975.

Rufus Burr Smith, professor of economics, Rollins College, Apr. 30, 1976.

## *Promotions*

Peter Asch: professor of economics, Rutgers–The State University, July 1, 1976.

Bennett D. Baack: associate professor of economics, Ohio State University.

Ernst Baltensperger: professor of economics, Ohio State University.

D. L. Brito: professor of economics, Ohio State University.

H. S. Burness: associate professor of economics, University of Kentucky, Sept. 1, 1976.

Robert M. Coen: professor of economics, Northwestern University.

Richard S. Cowan, Jr.: professor of business administration and economics, Waynesburg College, Sept. 1975.

Gerald S. Goldstein: associate professor of economics, Northwestern University.

Yukon Huang: associate professor, department of economics, University of Virginia, Sept. 1, 1976.

James E. Jonish: professor of economics, Texas Tech University, Apr. 1976.

Peter A. Koop: associate professor of economics, department of social sciences, King's College, July 1, 1976.

Conway L. Lackman: associate professor of economics, Rutgers–The State University, July 1, 1976.

Alton D. Law: professor of economics, Western Maryland College, Sept. 1, 1976.

John O. Ledyard: professor of economics, Northwestern University.

Dale T. Mortensen: professor of economics, Northwestern University.

Van Doorn Ooms: professor of economics, Swarthmore College, Mar. 1976.

Frederick W. Parkhurst, Jr.: professor of economics, Guilford College, June 1, 1976.

Hugh T. Rockoff: associate professor of economics, Rutgers–The State University, July 1, 1976.

Joseph J. Seneca: professor of economics, Rutgers–The State University, July 1, 1976.

Robert C. Stuart: professor, department of economics, Rutgers–The State University, July 1, 1976.

Bernard L. Weinstein: associate professor of economics and political economy, University of Texas at Dallas, Sept. 1, 1976.

## *Administrative Appointments*

Marcus Alexis: chairman, department of economics, Northwestern University, Sept. 1976.

Robert L. Avinger: dean, Center for Honor Studies Davidson College, July 1, 1976.

Phillippe J. Crabbé: chairman, department of economics, University of Ottawa, July 1, 1975.

Charles W. Hultman: interim chairman, department of economics, University of Kentucky, Jan. 1, 1976.

Peter A. Koop: chairman, department of social sciences, King's College, July 1, 1976.

John L. Madden: associate dean, College of Business

and economics, University of Kentucky, Sept. 1, 1975.

Frederick W. Parkhurst, Jr.: chairman, department of economics, Guilford College, June 1, 1976.

*Appointments*

John A. Andrulis: assistant professor, department of economics, Rutgers–The State University, July 1, 1976.

Charles W. Bausell, Jr., University of Maryland: assistant professor, department of economics, Kent State University, Sept. 1976.

Alberto Bensión: research associate, National Bureau of Economic Research, Apr. 23, 1976.

Bettina Berch, Barnard College: assistant professor, department of economics, Williams College, July 1, 1976.

Joel Bergsman, Congressional Budget Office: International Bank for Reconstruction and Development, May 3, 1976.

Theodore C. Bergstrom: visiting scholar, National Bureau of Economic Research, summer 1976.

Ralph Bradburd, State University of New York-Albany: assistant professor, department of economics, Williams College, July 1, 1976.

Ivan E. Brick: lecturer, department of economics, Rutgers–The State University, July 1, 1976.

Frank A. Camm, Jr., University of Chicago: associate economist, economics department, The Rand Corporation, Sept. 1976.

Stephen W. Chapel, Federal Energy Administration: senior economist, management sciences department, The Rand Corporation, June 1976.

Dov Chernichovsky: visiting scholar, National Bureau of Economic Research, summer 1976.

Douglas Coate: visiting scholar, National Bureau of Economic Research, summer 1976.

Paul H. Cootner: senior research associate, National Bureau of Economic Research, June 1976.

Vittorio Corbo: research associate, National Bureau of Economic Research, Apr. 23, 1976.

Bruce R. Dalgaard, University of Illinois: assistant professor of economics, Lehigh University, July 1, 1976.

John Dennis: research associate, National Bureau of Economic Research, Apr. 23, 1976.

Robert P. Flood: acting assistant professor, department of economics, University of Virginia, Sept. 1976.

Peter M. Garber: acting assistant professor, department of economics, University of Virginia, Sept. 1976.

David M. Gay: research associate, National Bureau of Economic Research, Apr. 23, 1976.

Fred Goldman: visiting scholar, National Bureau of Economic Research, summer 1976.

Daniel A. Graham: visiting research fellow, National Bureau of Economic Research, Sept. 1976.

William H. Greene, University of Wisconsin: assistant professor of economics, Cornell University, Sept. 1976.

Claudio Haddad: research associate, National Bureau of Economic Research, Apr. 23, 1976.

Eric A. Hanushek: research associate, National Bureau of Economic Research, June 1976.

Susan Hegsted, University of Utah: assistant professor of economics, Loyola University of Chicago, Sept. 1, 1976.

J. Stephen Henderson, University of Wisconsin: assistant professor, department of economics, Ohio State University.

John L. Hilley, Princeton University: assistant professor of economics, Lehigh University, Sept. 1, 1976.

Cheng Hsiao: visiting research fellow, National Bureau of Economic Research, Sept. 1976.

Douglas Joines, University of Chicago: lecturer of economics, Loyola University of Chicago, Feb. 1, 1976.

Arthur E. King, Ohio State University: assistant professor of economics, Lehigh University, Sept. 1, 1976.

Anne O. Krueger: senior research staff, National Bureau of Economic Research, Apr. 23, 1976.

I. David Lane: senior research associate, National Bureau of Economic Research, June 1976.

Jeffrey E. Levin: assistant professor, department of economics, University of Kentucky, Sept. 1, 1976.

James D. Libbin: research associate, department of economics, Iowa State University, Dec. 1, 1975.

Glenn C. Loury, Massachusetts Institute of Technology: assistant professor, department of economics, Northwestern University, Sept. 1976.

Stephen A. McCafferty, Brown University: assistant professor, department of economics, Ohio State University.

J. Huston McCulloch: visiting research fellow, National Bureau of Economic Research, Sept. 1976.

James M. McGrann: assistant professor, department of economics, Iowa State University, Jan. 1, 1976.

Roy E. Marsten: research associate, National Bureau of Economic Research, Apr. 23, 1976.

Stephen T. Marston, The Brookings Institution: assistant professor of economics, Cornell University, Sept. 1976.

Patricio Meller: research associate, National Bureau of Economic Research, Apr. 23, 1976.

Marc A. Miles: assistant professor, department of economics, Rutgers–The State University, July 1, 1976.

Jay B. Morrison: senior research associate, National Bureau of Economic Research, June 1976.

Leopold Mureithi: research associate, National Bureau of Economic Research, Apr. 23, 1976.

Mark Pitt: research associate, National Bureau of Economic Research, Apr. 23, 1976.

Richard A. Posner: senior research staff, National Bureau of Economic Research, Apr. 23, 1976.

David H. Pyle: senior research associate, National Bureau of Economic Research, June 1976.

Barr Rosenberg: senior research associate, National Bureau of Economic Research, June 1976.

Jeffrey I. Rubin: assistant professor, department of economics, Rutgers–The State University, July 1, 1976.

Frederic M. Scherer, Federal Trade Commission: professor, department of economics, Northwestern University, Sept. 1976.

T. Paul Schultz: research associate, National Bureau of Economic Research, Apr. 23, 1976.

Robert L. Schweitzer, University of North Carolina at Chapel Hill: assistant professor, department of economics, Kent State University, Sept. 1976.

Leslie Seplaki, Rutgers–The State University: visiting scholar, Harvard Law School, 1976–77.

Linda J. Shaffer, Northwestern University: assis-

tant professor of economics, University of California, Davis, July 1976.

William F. Sharpe: senior research associate, National Bureau of Economic Research, June 1976.

Steven M. Sheffrin, Massachusetts Institute of Technology: assistant professor of economics, University of California, Davis, July 1976.

Gordon R. Storholm: assistant professor, department of economics, Rutgers–The State University, July 1, 1976.

Carl Van Duyne, Stanford University: assistant professor, department of economics, Williams College, July 1, 1976.

W. Kip Viscusi, Harvard University: assistant professor, department of economics, Northwestern University, Sept. 1976.

Harvey Zabinsky: assistant professor, department of economics, University of Kentucky, Sept. 1, 1976.

### *Leaves for Special Appointments*

Christopher Garbcz, University of Missouri, Rolla: economist, Regional Impact Division, Federal Energy Administration, Washington, June 1, 1976–May 31, 1977.

Kurt Hausafus, Oberlin College: vice president and chief economist, Chicago Mercantile Exchange, June 1, 1976.

W. Whitney Hicks, University of Missouri, Columbia: Agency for International Development, Washington, Jan. 5, 1976.

Karel Holbik: chief, Section for Development of Financial Institutions, United Nations, New York Jan. 1976.

Mukul K. Majumdar, Cornell University: Ford Foundation rotation professorship, University of California, Berkeley; Guggenheim Fellowship, Sept. 1976.

### *Resignations*

Christopher C. Barnekov, Jr., Ohio State University, Sept. 1976.

Erwin A. Blackstone, Cornell University: Temple University, Sept. 1976.

Warren L. Coats, Jr., University of Virginia: International Monetary Fund, Jan. 15, 1976.

Stuart I. Greenbaum, University of Kentucky: Northwestern University, July 1, 1976.

Joseph A. McKinney, University of Virginia: Baylor University, May 1976.

Donald L. Martin, University of Virginia: University of Miami, May 1976.

Dennis C. Mueller, Cornell University: International Institute of Management, Berlin, Sept. 1976.

John A. Otte, Iowa State University, Feb. 4, 1976.

Pierre Pestieau, Cornell University: University of Liege, Sept. 1976.

### *Miscellaneous*

William G. Dewald: editor, *Journal of Money, Credit, and Banking.*

Peter O. Steiner, University of Michigan: president, American Association of University Professors, June 1976.

Roy Weinstein: vice president, National Economic Research Associates, Mar. 1, 1976.

## NOTE TO DEPARTMENTAL SECRETARIES AND EXECUTIVE OFFICERS

When sending information to the *Review* for inclusion in the Notes Section, please use the following style:

A. Please use the following categories:

1—Deaths
2—Retirements
3—Foreign Scholars (visiting the USA or Canada)
4—Promotions
5—Administrative Appointments
6—New Appointments
7—Leaves for Special Appointments (NOT Sabbaticals)
8—Resignations
9—Miscellaneous

B. Please give the name of the individual (SMITH, John W.), his present place of employment or enrollment: his new title (if any), and the date at which the change will occur.

C. Type each item on a separate 3x5 card and please do not send public relations releases.

D. The closing dates for each issue are as follows: *March*, November 1; *June*, February 1; *September*, May 1; *December*, August 1.

This announcement supersedes and replaces a letter which was sent annually from the managing editor's office. All items and information should be sent to the Assistant Editor, *American Economic Review*, Box Q, Brown University, Providence, Rhode Island 02912.

# ESSAYS ON HAYEK

**Edited by Fritz Machlup. Foreword by Milton Friedman**

Friedrich A. Hayek, Nobel Laureate in Economic Science 1974

Focusing on one of this century's most influential economic and political thinkers, this book brings together seven original essays presented at a special meeting of the Mont Pelerin Society, plus additional commentary about Hayek and his work.

WILLIAM F. BUCKLEY, Jr.—*The Road to Serfdom: The Intellectuals and Socialism*
GOTTFRIED DIETZE—*Hayek on the Rule of Law*
RONALD MAX HARTWELL—*Capitalism and the Historians*
SHIRLEY ROBIN LETWIN—*The Achievements of Friedrich A. Hayek*
FRITZ MACHLUP—*Hayek's Contribution to Economics*
GEORGE C. ROCHE III—*The Relevance of Friedrich A. Hayek*
ARTHUR SHENFIELD—*Scientism and the Study of Society*

Clothbound, $10.00

**NYU Press**
Washington Square, New York 10003

# Insights: International Economics

**Foreign Trade and International Development**
**Anindya K. Bhattacharya,** Adelphi University
The author considers such issues as international and domestic income distribution, long-term trading trends, and preferential tariff treatment in discussing the progress of less-developed nations. 128pp. $13.50

**The New Internationalism:** Strategy and Initiatives for U.S. Foreign Economic Policy
**Wilbur F. Monroe**
The author conducts an ambitious, knowledgeable investigation, by blocs of countries, into the make-up of a comprehensive foreign economic policy. 256pp. $16.00

**China's Foreign Aid:** An Instrument of Peking's Foreign Policy
**John F. Copper,** University of Maryland
Here is an intriguing analysis of China's economic aid program during the past twenty years and its relationship to foreign-policy goals. 224pp. $17.50

**Latin American Inflation**
**Susan Wachter,** University of Pennsylvania
Unlike other works dealing with inflationary problems of Latin America, Wachter reviews the monetarist-structuralist debate, reformulating the structuralist model for testing. The author subjects data from Chile, Brazil, Argentina and Mexico to rigorous analysis. In Press Fall 1976

HEATH D. C. HEATH AND COMPANY
125 Spring Street
Lexington, Massachusetts 02173

**Lexington Books**

# *ANNOUNCING*

# A SECOND JOB PLACEMENT MEETING

**O'HARE HILTON**

**CHICAGO, ILLINOIS**

**January 7, 8, & 9**

As the annual meeting is being held at an early point in the academic year (September 16-18, 1976), it may not be possible to complete job placement activity through the office of the Employment Registry and Center. Accordingly, the American Economic Association will hold a special supplementary placement meeting on January 7, 8, and 9, 1976 at the O'Hare Hilton in Chicago, Illinois. The O'Hare Hilton is located within the airport complex and is within easy walking distance of the terminals.

---

Instructions regarding application submission, job listing submission, and hotel reservations will be available in the October issue of *JOB OPENINGS FOR ECONOMISTS*, the December issue of the *AMERICAN ECONOMIC REVIEW*, and the Secretary's Office (after the September Annual Meeting). In addition, information for interested employers and applicants will be distributed to all Department Chairpersons and to nonacademic employers who have listed openings in *JOE*.